METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Chemoinformatics and Computational Chemical Biology

Edited by

Jürgen Bajorath

Department of Life Science Informatics, Rheinische Friedrich-Wilhelms-Universität, B-IT, LIMES
Dahlmannstr. 2, 53113 Bonn, Germany

Humana Press

Editor
Jürgen Bajorath
Department of Life Science Informatics
Rheinische Friedrich-Wilhelms-Universität
B-IT, LIMES
Dahlmannstr. 2
53113 Bonn
Germany

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-60761-838-6 e-ISBN 978-1-60761-839-3
DOI 10.1007/978-1-60761-839-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010936182

© Springer Science+Business Media, LLC 2011
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Humana Press, c/o Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights. While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Humana press is a part of Springer Science+Business Media (www.springer.com)

Preface

In 2004, vol. 275 of the *Methods in Molecular Biology*™ series was published. This book, entitled *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, presented an array of different chemoinformatics methodologies. Now, 6 years later, a second volume focusing on chemoinformatics is introduced with the title *Chemoinformatics and Computational Chemical Biology*. Besides new focal points, there is a link between these volumes because eight of the lead authors who contributed in 2004 also contribute to the new book.

Over the past years, the chemoinformatics field has further evolved and new application areas have opened up, one of which is highlighted in the new book. Other developing application areas for chemoinformatics approaches could have also been emphasized, but chemical biology, the study of biological functions and systems using small molecules, seemed particularly appropriate, given that this field is at least distantly related to pharmaceutical research, which has been one of the origins of chemoinformatics (and also a major focal point of the 2004 volume). Topics of interest in chemical biology that can be addressed with the aid of chemoinformatics methodologies include, among others, system-directed approaches using small molecules, the design of target-focused compound libraries, the study of molecular selectivity, or the systematic analysis of target–ligand interactions, all of which are discussed in the book.

Currently, both long-established computational approaches and new methodologies are considered part of the chemoinformatics spectrum, and the book also aims to reflect this situation. Thus, in addition to topics relevant for chemical biology, which are mostly discussed in the last third of the book, mainstays of chemoinformatics are covered including similarity methods, machine learning, probabilistic approaches, and fragment-based methods. Other contributions concentrate on structure–activity relationships and underlying activity landscapes, pharmacophore concepts, de novo ligand design, and chemical reaction modeling. Many contributions discuss issues related to virtual compound screening. Two chapters even go beyond the current chemoinformatics spectrum by describing knowledge-based modeling of G protein-coupled receptor structures and computational design of siRNA libraries. The book begins with a detailed introduction into the chemoinformatics field and its development and ends with a discussion of statistical standards for the evaluation of virtual screening calculations, a topic of general relevance.

This book has brought together a group of leading investigators, both from academia and industry, who have helped to shape the chemoinformatics field. Eighteen chapters were solicited from different investigators to cover various methodological aspects of chemoinformatics. As the book evolved, four contributions from our group were added to complement and further expand selected research areas. More than half of the 22 chapters have review-type character; the others describe an individual method or a class of methods. The sequence of chapters follows a logical flow, to the extent possible, and chapters having thematic connections are presented back-to-back.

It is hoped that this compendium of articles will be of interest to experts, but also newcomers to this exciting field. I am very grateful to our authors whose contributions have made this book possible.

Bonn, Germany, January 18, 2010 ***Jürgen Bajorath***

Contents

Preface ..... *V*
Contributors ..... *IX*

1 Some Trends in Chem(o)informatics 1
Wendy A. Warr

2 Molecular Similarity Measures 39
Gerald M. Maggiora and Veerabahu Shanmugasundaram

3 The Ups and Downs of Structure–Activity Landscapes 101
Rajarshi Guha

4 Computational Analysis of Activity and Selectivity Cliffs 119
Lisa Peltason and Jürgen Bajorath

5 Similarity Searching Using 2D Structural Fingerprints 133
Peter Willett

6 Predicting the Performance of Fingerprint Similarity Searching 159
Martin Vogt and Jürgen Bajorath

7 Bayesian Methods in Virtual Screening and Chemical Biology 175
Andreas Bender

8 Reduced Graphs and Their Applications in Chemoinformatics 197
Kristian Birchall and Valerie J. Gillet

9 Fragment Descriptors in Structure–Property Modeling and Virtual Screening 213
Alexandre Varnek

10 The Scaffold Tree: An Efficient Navigation in the Scaffold Universe 245
Peter Ertl, Ansgar Schuffenhauer, and Steffen Renner

11 Pharmacophore-Based Virtual Screening 261
Dragos Horvath

12 De Novo Drug Design 299
Markus Hartenfeller and Gisbert Schneider

13 Classification of Chemical Reactions and Chemoinformatics Processing of Enzymatic Transformations 325
Diogo A.R.S. Latino and João Aires-de-Sousa

14 Informatics Approach to the Rational Design of siRNA Libraries 341
Jerry O. Ebalunode, Charles Jagun, and Weifan Zheng

15 Beyond Rhodopsin: G Protein-Coupled Receptor Structure and Modeling Incorporating the β2-adrenergic and Adenosine A_{2A} Crystal Structures 359
Andrew J. Tebben and Dora M. Schnur

16 Methods for Combinatorial and Parallel Library Design 387
Dora M. Schnur, Brett R. Beno, Andrew J. Tebben, and Cullen Cavallaro

17 The Interweaving of Cheminformatics and HTS 435
Anne Kümmel and Christian N. Parker

18 Computational Systems Chemical Biology 459
Tudor I. Oprea, Elebeoba E. May, Andrei Leitão, and Alexander Tropsha

19 Ligand-Based Approaches to In Silico Pharmacology 489
David Vidal, Ricard Garcia-Serna, and Jordi Mestres

20 Molecular Test Systems for Computational Selectivity Studies and Systematic Analysis of Compound Selectivity Profiles 503
Dagmar Stumpfe, Eugen Lounkine, and Jürgen Bajorath

21 Application of Support Vector Machine-Based Ranking Strategies to Search for Target-Selective Compounds. 517
Anne Mai Wassermann, Hanna Geppert, and Jürgen Bajorath

22 What Do We Know?: Simple Statistical Techniques that Help 531
Anthony Nicholls

Index 583

Contributors

João Aires-de-Sousa · *CQFB and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal*

Jürgen Bajorath · *Department of Life Science Informatics, B-IT, LIMES, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Andreas Bender · *Faculty of Science, Gorlaeus Laboratories, Center for Drug Research, Medicinal Chemistry, Universiteit Leiden/Amsterdam, Leiden, The Netherlands*

Brett R. Beno · *Computer Aided Drug Design, Pharmaceutical Research Institute, Bristol-Myers Squibb Company, Princeton, NJ, USA*

Kristian Birchall · *Department of Chemistry, University of Sheffield, Sheffield, UK*

Cullen Cavallaro · *Computer Aided Drug Design, Pharmaceutical Research Institute, Bristol-Myers Squibb Company, Princeton, NJ, USA*

Jerry O. Ebalunode · *Department of Pharmaceutical Sciences, BRITE Institute, North Carolina Central University, Durham, NC, USA*

Peter Ertl · *Novartis Institutes for BioMedical Research, Basel, Switzerland*

Ricard Garcia-Serna · *Chemotargets SL and Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain*

Hanna Geppert · *Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Valerie J. Gillet · *Department of Information Studies, University of Sheffield, Sheffield, UK*

Rajarshi Guha · *NIH Chemical Genomics Center, Rockville, MD, USA*

Markus Hartenfeller · *Institute of Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe University, Frankfurt, Germany*

Dragos Horvath · *Laboratoire d'InfoChime, UMR 7177 Université de Strasbourg – CNRSInstitut de Chimie, Strasbourg, France*

Charles Jagun · *Department of Pharmaceutical Sciences, BRITE Institute, North Carolina Central University, Durham, NC, USA*

Anne Kümmel · *Novartis Institutes for BioMedical Research, Basel, Switzerland*

Diogo A.R.S. Latino · *CQFB and REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal*

Andrei Leitão · *Department of Biochemistry and Molecular Biology, School of Medicine, University of New Mexico, Albuquerque, NM, USA*

Eugen Lounkine · *Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*

Gerald M. Maggiora · *Department of Pharmacology & Toxicology, College of Pharmacy, University of Arizona, Tucson, AZ, USA*

Elebeoba E. May · *Sandia National Laboratories, Complex Systems and Discrete Mathematics, Albuquerque, NM, USA*

Jordi Mestres · *Chemotargets SL and Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain*
Anthony Nicholls · *OpenEye Scientific Software, Santa Fe, NM, USA*
Tudor I. Oprea · *Department of Biochemistry and Molecular Biology, School of Medicine, University of New Mexico, Albuquerque, NM, USA*
Christian N. Parker · *Novartis Institutes for BioMedical Research, Basel, Switzerland*
Lisa Peltason · *Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*
Steffen Renner · *Novartis Institutes for BioMedical Research, Basel, Switzerland*
Gisbert Schneider · *Institute of Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe University, Frankfurt, GermanyInstitute of Pharmaceutical Sciences, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland*
Dora M. Schnur · *Computer Aided Drug Design, Pharmaceutical Research Institute, Bristol-Myers Squibb Company, Princeton, NJ, USA*
Ansgar Schuffenhauer · *Novartis Institutes for BioMedical Research, Basel, Switzerland*
Veerabahu Shanmugasundaram · *Anti-Bacterials Computational Chemistry, Department of Structural Biology, WorldWide Medicinal Chemistry, Pfizer Global Research & Development, Groton, CT, USA*
Dagmar Stumpfe · *Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*
Andrew J. Tebben · *Computer Aided Drug Design, Pharmaceutical Research Institute, Bristol-Myers Squibb Company, Princeton, NJ, USA*
Alexander Tropsha · *Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA*
Alexandre Varnek · *Laboratory of Chemoinformatics, UMR 7177 CNRS, University of Strasbourg, Strasbourg, France*
David Vidal · *Chemotargets SL and Chemogenomics Laboratory, Research Unit on Biomedical Informatics, Institut Municipal d'Investigació Mèdica and Universitat Pompeu Fabra, Barcelona, Catalonia, Spain*
Martin Vogt · *Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*
Wendy A. Warr · *Wendy Warr & Associates, Holmes Chapel, Cheshire, UK*
Anne Mai Wassermann · *Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany*
Peter Willett · *Department of Information Studies, The University of Sheffield, Sheffield, UK*
Weifan Zheng · *Department of Pharmaceutical Sciences, BRITE Institute, North Carolina Central University, Durham, NC, USA*

Chapter 1

Some Trends in Chem(o)informatics

Wendy A. Warr

Abstract

This introductory chapter gives a brief overview of the history of cheminformatics, and then summarizes some recent trends in computing, cultures, open systems, chemical structure representation, docking, de novo design, fragment-based drug design, molecular similarity, quantitative structure–activity relationships (QSAR), metabolite prediction, the use of phamacophores in drug discovery, data reduction and visualization, and text mining. The aim is to set the scene for the more detailed exposition of these topics in the later chapters.

Key words: History of cheminformatics, Cheminformatics, Chemical structures, 2D searching, 3D searching, Similarity, Protein–ligand docking, Virtual high throughput screening, De novo design, Fragment-based drug design, Ligand-based drug design, QSAR, Computing infrastructure, Pharmaceutical industry, Open systems, Metabolite prediction, Pharmacophore, Visualization, Text mining

1. Introduction

Despite the fact that chem(o)informatics started to emerge as a distinct discipline in the late 1990s, its practitioners have still not alighted on a title for it [1, 2]: the terms "cheminformatics," "chem(o)informatics," "chemical informatics," and "chemical information" have all been used, but "cheminformatics" is the most commonly used name [3]. The *Journal of Chemical Information and Modeling* (formerly the *Journal of Chemical Information and Computer Sciences*), the core journal for the subject [2], still uses the term "chemical information" despite the fact that until recently chemical information was a discipline more associated with databases and "research information" [4, 5].

Not only is there no agreement on the name of the discipline, but also there is no agreed definition of what is involved [1, 2].

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_1, © Springer Science+Business Media, LLC 2011

Despite all this, practitioners do belong to a certain community, and one that is truly international in nature. It is possible to distinguish the learned journals that publish papers in the field [2, 6]. The *Journal of Chemical Information and Modeling* is the core journal but many significant papers are published in journals whose principal focus is molecular modeling or quantitative structure–activity relationships (QSAR) or more general aspects of chemistry [2]. Typical specialized journals are the *Journal of Medicinal Chemistry*, the *Journal of Computer-Aided Molecular Design*, the *Journal of Molecular Graphics and Modelling*, and *QSAR & Combinatorial Science*, but cheminformatics papers also appear in more broadly based journals such as *Drug Discovery Today* and *Current Opinion in Drug Discovery and Development*. The textbooks most commonly used in cheminformatics courses are those by Leach and Gillet [7] and Gasteiger and Engel [8]. Books edited by Bajorath [9] and by Oprea [10] are also recommended. Schneider and Baringhaus have produced a more recent text book [11].

The current book, *Chemoinformatics and Computational Chemical Biology*, is a successor to Bajorath's 2004 book *Chemoinformatics: concepts, methods, and tools for drug discovery* [9]. Its chapters cover such a wide range of topics that it is not possible in this introductory article to give a detailed analysis of trends in each of the fields. Instead, some general trends will be highlighted, and illustrated with a selected and by no means comprehensive set of examples. First, there is a very brief history of selected fields in cheminformatics up to about the year 2000, and after that newer developments in each field are considered in separate sections.

2. History

Several useful histories have been published recently [12–16]. Chemical structure handling is a mature technology [17–20]. Chemical databases, structure and substructure searching in 2D [17–21], reaction retrieval [17, 22], generation of 3D structures from 2D structures [23–26], 3D searching [27–34], similarity search of 2D or 3D structures [32, 35], and retrieval of generic ("Markush") structures [36, 37] are well documented.

Reaction systems can be subdivided into reaction retrieval systems (such as REACCS and its successors from Symyx Technologies, CASREACT from Chemical Abstracts Service, and Reaxys from Elsevier) and synthetic analysis programs [38]. Ott [22] further divides the latter class into synthesis design programs (such as LHASA, which relies on a knowledge base) and WODCA (now THERESA) [39], which performs retrosynthesis in a logic-oriented fashion); reaction prediction programs, such as Ugi's

IGOR [40, 41], Gasteiger's EROS [42], and Jorgensen's CAMEO [43]; and mechanism elucidation programs.

Programs for the generation of 3D structures from 2D structures [23–26, 44, 45] spurred on development of 3D structure methods because at that time only a limited number of experimentally determined 3D structures were available in databases. The history of crystallographic databases goes back to the early 1970s, but it has taken many years for them to grow. The Cambridge Structural Database [46–49] is the world repository of small molecule crystal structures. By January 2009 it contained 469,611 structures. The Protein Data Bank (PDB) began as a grassroots effort in 1971. It has grown from a small archive containing a dozen structures to a major international resource for structural biology containing more than 40,000 entries [50]. It contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.

If the structure of a drug target is known (e.g. is in the PDB) it can be used in structure-based drug design. In computer-aided drug design four situations can arise. In the first case, if the receptor structure is unknown and there are no known ligands, computational techniques can be used to design collections of compounds with diverse structures for high throughput screening (HTS). In another case, the structure of the target or receptor is known, and the structure of the ligand is known. In this case protein-docking algorithms can be used to place candidate compounds within the active site of the target and rank-order them. In the third situation, the structure of the receptor is known, but the structure of the ligand is unknown; de novo design techniques can propose new ligands that are complementary to the active site.

In the final case, where the receptor structure is unknown, but there are ligands of known structure, ligand-based drug design is carried out. Similarity-based and machine learning-based "virtual screening" can be used, for example, or starting with a collection of molecules of known activity, the computational chemist can develop either a QSAR model or a 3D pharmacophore hypothesis that is converted into a search query. Scientists then use the search query to search a 3D database for structures that fit the hypothesis, or they use the QSAR model to predict activities of novel compounds.

Early programs for pharmacophore mapping (also called elucidation) [51] included DISCO [52], Catalyst [53, 54], and Genetic Algorithm Superimposition Program GASP [55–57]. Catalyst is still in regular use today. It has two components: HypoGen, a regression-like method for generating and optimizing hypotheses from 15–30 compounds with a range of potencies and HipHop for finding key features and producing alignments from a small set of potent compounds. Pharmacophore tools must allow for the fact that many compounds are flexible and can

assume multiple conformations. The exploration of multiple conformations (and often large numbers of them) can be tackled by generating and storing multiple representative conformations (as in Catalyst) or by exploring conformational space "on the fly."

GASP employs a genetic algorithm (GA) for the superimposition of sets of flexible molecules. Molecules are represented by a chromosome that encodes angles of rotation about flexible bonds and mappings between pharmacophore-like features in pairs of molecules. The molecule with the smallest number of features in the data set is used as a template, onto which the remaining molecules are fitted. The fitness function of the GA is a weighted combination of the number and the similarity of the features that have been overlaid in this way; the volume integral of the overlay; and the van der Waals energy of the molecular conformations defined by the torsion angles encoded in the chromosomes.

In ligand-based drug design, an alternative to the use of pharmacophores is QSAR. The QSAR method involves the conversion of molecular structures into mathematical descriptors that capture the properties of molecules that are relevant to the activity being modeled; selecting the best descriptors from a large set; mapping those descriptors onto the activities; and validating the model to determine how predictive it is and how well it can be extrapolated to molecules not in the training set used to generate the model [58–62]. Descriptors may be calculated from 2D or 3D structures. In the widely used Comparative Molecular Field Analysis (CoMFA) method [63], a 3D structure is surrounded by an array of grid points. At each point outside the molecule a probe atom or functional group is used to calculate steric, electrostatic, and sometimes, lipophilic fields at that point. One disadvantage is that an alignment rule is needed to superimpose molecules in the training set. A related method is comparative molecular similarity indices analysis (CoMSIA) [64]. FlexS [65] can be used to align molecules in 3D and prepare compounds for 3D QSAR analysis.

Deficiencies in absorption, distribution, excretion, metabolism (ADME) characteristics are the leading causes of attrition during drug development [66]. Prediction of toxicology is particularly difficult because of the variety of biological processes involved, but much research has been carried out in this field [67, 68]. Lipinski's "rule of five" [69] is a widely used rule of thumb to predict "druglikeness." Lipinski's rule says that, in general, an orally active drug has no more than one violation of the following criteria: not more than five hydrogen bond donors, not more than 10 hydrogen bond acceptors, a molecular weight under 500 Da, and an octanol–water partition coefficient (Clog*P*) of less than 5.

Widespread adoption of HTS and chemical synthesis technologies in the 1990s led to a data deluge. Combinatorial chemistry

allows very large numbers of chemical entities to be synthesized by condensing a small number of reagents together in all possible combinations. A "chemical library" is a set of mixtures or discrete compounds made by one combinatorial reaction. As an indication of the size of the chemical space covered by one library, Cramer [70] quotes the reaction of 4,145 commercially available diamines with R groups from 68,934 acylating reagents, cleanly displaceable halides, etc. (used twice), giving a library 2.0×10^{13} compounds. Compare this number with the 50 million known compounds in the CAS Registry. Cramer calculates that it would take 60,000 years of screening at the rate of 1 million per day, to test 2.0×10^{13} compounds. In addition, random screening has proved too expensive, its hit rate is low, false positives may be a problem, and expensive compounds are consumed. At the same time, developments in hardware and software during the 1990s meant that larger amounts of 3D structural information could be processed and allowed new methodological approaches to computer-aided drug discovery [32, 71]. All of this proved fruitful for progress in computational chemistry.

Selection of those compounds most likely to be hits was of considerable interest. It has been said (in a statement often attributed to David Weininger) that there are 10^{180} possible compounds, 10^{18} likely drugs, 10^{7} known compounds, 10^{6} commercially available compounds, 10^{6} compounds in corporate databases, 10^{4} compounds in drug databases, 10^{3} commercial drugs, and 10^{2} profitable drugs. Early library design efforts [14, 35, 72–74] involved selecting diverse subsets by clustering, dissimilarity-based selection, partitioning/cell-based approaches, or optimization-based methods [75]. Initially, diverse or focused libraries were designed based on descriptors for the reagents, since this required less computation, but later it was shown that product-based design produces more diverse libraries [76]. Moreover, diversity (or similarity, with design focused on certain specific chemical series) is not the only criterion for compound selection. Other factors such as druglikeness and synthetic accessibility need to be considered. This need led to progress in the field of multiobjective library design [77, 78].

To be considered for further development, lead structures should be members of an established SAR series; and should have simple chemical features, amenable to chemical optimization, a favorable patent situation, and good ADME properties. By analyzing two distinct categories of leads, those that lack any therapeutic use (i.e. "pure" leads), and those that are marketed drugs themselves but have been altered to yield novel drugs, Oprea and colleagues [79, 80] have shown that the process of optimizing a lead into a drug results in more complex structures. Hann and co-workers have studied molecular complexity [81]

and shown that less complex molecules are more common starting points for the discovery of drugs. These studies of leadlikeness and druglikeness have contributed to a trend in designing libraries of less complex, leadlike molecules, to leave scope for the almost inevitable increase in size and complexity during the optimization process.

Libraries may be designed for biological screening purposes, but computers may also be used to predict the results of screening in a process called "virtual screening." Nowadays, the term "virtual high throughput screening" is sometimes equated with protein–ligand docking, but other methods for virtual HTS have been developed, including identifying drug-like structures [82, 83], 2D similarity [35, 84], pharmacophore elucidation, and the use of 3D pharmacophores in 3D database searching [27–30, 32, 71]. An algorithm for docking small molecules to receptors (later to become the DOCK program) was published by Kuntz et al. [85] as long ago as 1982. Other programs started to appear in the late 1990s [55–57, 86]. Many factors, including an increase in the power of computers, have made docking increasingly popular of late.

3. Computers and Computing Environments

It is likely that a man was put on the moon in 1969 using a computer less powerful than today's typical cell phone. The cheminformatics systems of the 1980s and 1990s were run on mainframes or "minis" less powerful than today's PC. The VAX 11/750 used to run the U.K. Chemical Database Service [87] in 1984 had a clock speed of 6 MHz, 2 Mb memory, 134 Mb fixed disk, and two 67 Mb exchangeable disk drives; by judicious sharing of peripherals, the cost was kept down to £100,000 (1984 price). Readers need only to look at the specification and cost of their own hardware to see the advances of the last 20 years. Nowadays, parallel code and grid computing are commonplace, and cyberinsfrastructure (e-science) has been established [88] as an enabling platform.

For applications needing more speed than the average CPU can provide, field programmable gateways (FPGAs), graphics processing units (GPUs), and even gaming devices such as Microsoft's Xbox gave been explored. For example, the so-called Lightning version of SimBioSys' eHiTS ligand docking software [89] has been run on the Sony PlayStation 3 (PS3) game console [90], or more specifically on a microprocessor architecture called the Cell Broadband Engine (Cell/B.E.) which powers the PS3. The Cell/B.E. enables the PS3 to speed up physics simulations so that

they can catch up with the 3D graphics rendering speeds of the system's GPUs. The IBM BladeCenter QS21 blade server is based on the same Cell/B.E. processor.

The emergence of the World Wide Web in 1992 brought about an information revolution. Now "Web 2.0" technologies [91–93] and the Semantic Web [94–97] are having an impact. "Web 2.0" is a badly defined term covering multiple collaborative technologies such as instant messaging, text chat, Internet forums, weblogs ("blogs"), wikis, Web feeds, and podcasts; social network services including guides, bookmarking, and citations; and virtual worlds such as Second Life [91]. Some of these technologies may seem to have little relevance to cheminformatics, but there are in fact some applications of interest, e.g. Pfizerpedia [98], the Pfizer wiki, use of which has reportedly increased exponentially.

In 2005, CAS introduced CAS Mobile for real-time interaction with CAS databases using wireless handheld devices. In 2009, the application ChemMobi was posted to the Apple App Store and can be downloaded, for free, to enable an iPhone to search both Symyx's Discovery Gate [99] and ChemSpider [100]. ChemMobi uses DiscoveryGate Web Service and the ChemSpider Web Service: Web Services are another feature of today's computing architectures. "Cloud computing" (information infrastructure, software, and services hosted on the Internet rather than on one's own computer) is in its infancy but is attracting much interest in the technology press.

Pipelining and workflow methods have long been used in bioinformatics but later started to impact cheminformatics [101]. The workflow paradigm is a generic mechanism to integrate different data resources, software applications and algorithms, Web services, and shared expertise. (There are minor differences between pipelining and workflow.) Such technologies allow a form of integration and data analysis that is not limited by the restrictive tables of a conventional database system. They enable scientists to construct their own research data processing networks (sometimes called "protocols") for scientific analytics and decision making by connecting various information resources and software applications together in an intuitive manner, without any programming. Therefore, software from a variety of vendors can be assembled into something that is the ideal workflow for the end user. These are, purportedly, easy-to-use systems for controlling the flow and analysis of data. In practice, certainly in chemical analyses, they are not generally used by novices: best use of the system can be made by allowing a computational chemist to set up the steps in a protocol then "publish" it for use by other scientists. KNIME and Pipeline Pilot are the systems most familiar to computational chemists but Kepler, Taverna, SOMA, and InforSense are also worthy of note [101].

4. Influence of Industry

Much of the research in this subject area is carried out in industry and the majority of the applications are written in industry [6], and in particular the pharmaceutical industry [102]. Thus, it is not surprising that a significant number of the authors in this book have, or have had, industrial affiliations. The pharmaceutical industry, however, currently faces unprecedented problems, including the increasing cost of R&D, the decreasing number of new chemical entities, patent expiry on "blockbuster" drugs between 2008 and 2013, and pressure to reduce drug prices. This has led to new strategies and increased interest in translational research, pharmacogenomics, biomarkers, and personalized medicine. Cheminformatics is likely to change in response to these changes.

Industry and academia have always collaborated in cheminformatics projects but a new trend is the availability in academia of biological assay data: the sort of data that used to be largely held in proprietary systems in the pharmaceutical industry. Data from the Molecular Libraries Screening Centers in the United States are available through PubChem [103, 104]. The National Institutes of Health (NIH) are also launching the Therapeutics for Rare and Neglected Diseases (TRND) program [105] which creates a drug development pipeline within the NIH and is specifically intended to stimulate research collaborations with academic scientists working on rare illnesses. The cultures of bioinformatics and cheminformatics have tended to differ when it comes to open applications and sharing of data: cheminformaticians are the more likely to use proprietary software and data.

5. Open Systems

The pharmaceutical industry in general has not been keen to embrace open source software, although some individuals are enthusiastic [106, 107]. Limited attempts have been made at collaboration (the commercially available organic chemicals alliance in the 1980s was a success in some respects, and LHASA Limited continues to this day), but the current trend to open systems has led to renewed interest in collaborating in areas which do not give competitive advantage [108, 109].

The Pistoia Alliance [108] started with an initial meeting (in Pistoia, Italy) where proponents at GlaxoSmithKline, AstraZeneca, Pfizer, and Novartis outlined similar challenges and frustrations in the IT and informatics sector of discovery. The advent

of Web Services and Web 2.0 allows for decoupling of proprietary data from technology. A service orientated approach allows for these types of pre-competitive discussions. The primary purpose of the Pistoia Alliance is to streamline non-competitive elements of the pharmaceutical drug discovery workflow by the specification of common business terms, relationships, and processes. There is a vast amount of duplication, conversion, and testing that could be reduced if a common foundation of data standards, ontologies, and Web Services could be promoted and ideally agreed within a non-proprietary and non-competitive framework. This would allow interoperability between a traditionally diverse set of technologies to benefit the healthcare sector. The Pistoia Alliance was officially launched in February 2009. Additional members include ChemITment, ChemAxon, Accelrys, Edge Consultancy, BioXPR, GGA, Lundbeck, Bristol Myers Squibb, Roche, KNIME, Rescentris, and DeltaSoft.

A case study is the LHASA Web Service. Before Pistoia involvement, LHASA's DEREK [110] software did not have a Web Service; it had only a Windows-based API. Each company had created or adapted its own LHASA interface, which was subject to change as LHASA updated the product. Each company had done much the same interfacing. There was no consistent approach to a Web Service. After Pistoia involvement, there is a single LHASA DEREK Web Service available to all customers.

Similar trends are occurring in bioinformatics [109]. In cheminformatics, there is now considerable interest in "open source, open standards, open data," and "reusable chemistry" [91, 92, 94, 95]. Large, chemical structure databases with physical property and activity data such as ZINC [111], PubChem [103], ChemSpider [100], eMolecules [112], and the NIH/CADD Chemical Structure Lookup Service [113] have become freely available [114]. These databases contain millions of molecules but even larger databases of virtual molecules have been produced [115, 116]. The Collaborative Drug Discovery [117] portal enables scientists to archive, mine, and collaborate around preclinical chemical and biological drug discovery data through a Web-based interface. Free exchange of chemical structures on the Web has become easier with the emergence of open standards for identification.

6. Chemical Structure Representation

The Morgan algorithm [118] underpins many of the systems in use today, and is the basis of the CAS REGISTRY database. It identifies atoms based on an extended connectivity value; the atom with the highest value becomes the first atom in the

name, and its neighbors are then listed in descending order. Ties are resolved based on additional parameters, for example, bond order and atomic number. The original Morgan algorithm did not handle stereochemistry; the Stereochemically Extended Morgan Algorithm (SEMA) was developed to handle stereoisomers [119]. The Newly Enhanced Morgan Algorithm (NEMA) [120] produces a unique name and key for a wider range of structures than SEMA. It extends perception to non-tetrahedral stereogenic centers, it supports both 2D and 3D stereochemistry perception, and it does not have an atom limit. The CAS Registry system, SEMA and NEMA are proprietary.

Systems such as ChemSpider use the International Union of Pure and Applied Chemistry (IUPAC) International Identifier (InChI) to register chemical structures. IUPAC developed InChI as a freely available, non-proprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations and unambiguous identification of chemical substances. IUPAC decided to tackle this problem because the increasing complexity of molecular structures was making conventional naming procedures inconvenient, and because there was no suitable, openly available electronic format for exchanging chemical structure information over the Internet. The goal of the IUPAC International Chemical Identifier (InChI) is to provide a unique string representing a chemical substance of known structure, independent of specific depiction, derived from conventional connection table, freely available, and extensible [121–123].

The InChI project was initially undertaken by IUPAC with the cooperation of National Institute for Standards and Technology (NIST). Steps 1–3 of the canonical numbering for InChI are done using an algorithm modified from that of McKay [124]. In 2009, a standard version of InChI and the InChIKey were released. InChIKey is a fixed length condensed digital representation of the identifier which facilitates Web searching, previously complicated by unpredictable breaking of InChI character strings by search engines. It also allows development of a Web-based InChI lookup service; permits an InChI representation to be stored in fixed length fields; and makes chemical structure database indexing easier. InChI has been used in chemical enhancement of the Semantic Web [94].

Like InChI, the SMILES language [125, 126] allows a canonical serialization of molecular structure. However, SMILES is proprietary and unlike InChI, it is not an open project. This has led to the use of different generation algorithms, and thus, different SMILES versions of the same compound have been found. InChI is not a registry system such as that of CAS; it does not depend on the existence of a database of unique substance records

to establish the next available sequence number for any new chemical substance being assigned an InChI.

Having looked at recent trends in the basics of cheminformatics (hardware, infrastructures, and cultural issues) we will now consider drug design technologies one by one, starting with protein–ligand docking, an approach which has proved increasingly popular in recent years.

7. Docking

Receptor–ligand docking [127] is a computational procedure that predicts the binding mode and affinity of a ligand to a target receptor. In this method, each ligand in a database is docked into the active site of the receptor. Docking programs require an algorithm that can explore a very large number of potential docking conformations and orientations for each ligand. These programs generate a series of 3D models that predict the way in which each small molecule will bind to the targeted receptor. Docking programs also include a scoring function that quantitatively ranks the ligands according to their binding affinity for the receptor. Although scoring functions are meant to determine which compounds are more likely have a higher affinity for the target molecule, in practice, these functions do not always accurately rank ligands.

More than 60 docking programs (e.g., Autodock, DOCK, eHiTS, FlexX, FLOG, FRED, Glide, GOLD, Hammerhead, ICM, Surflex-Dock) and more than 30 scoring functions [128] have been recorded. Many different docking methods are available including fast shape matching, distance geometry, genetic algorithms, simulated annealing, incremental construction, and tabu search. Scoring has been carried out with force fields, empirical schemes, potentials of mean force, linear interaction energies, or other molecular dynamics.

In theory, docking methods provide the most detailed description of all virtual screening techniques. Ideally, ligand–receptor docking methods describe how a compound will interact with a target receptor, what contacts it will make within the active site, and what binding affinity it will have for the receptor. Although pharmacophore-based searching can select compounds possessing approximately the desired 3D characteristics, this approach does not describe how well matched the molecule and the receptor are in terms of shape or other properties and does not predict binding affinity or rank the selection in any way. Also, while statistical selection methods are excellent at finding compounds with structural similarities to known ligands, docking methods are better suited to finding novel structural types.

Docking, however, is such a computationally demanding exercise that, in practice, its limitations or approximations are inevitable. For example, although many current docking programs account for some degree of ligand flexibility, sampling [128] of the many possible conformations of each ligand in a database must be limited if the search is to be completed in a reasonable length of time. Also, when a ligand binds to its receptor, the receptor may also undergo conformational change. Docking programs that use rigid receptor models do not account for these changes.

Another important issue for researchers using docking as a virtual screening strategy is the accuracy of the predicted binding modes and affinities. Studies have shown that while some docking programs can reproduce crystal structures well, they generate scores that do not correlate with measured IC_{50} values. Of particular concern is the ability of current scoring functions to rank ligands correctly. To dock all the molecules contained in very large database, a scoring function must be simple, very fast, trained across a wide variety of proteins, and derived from a physically reasonable equation. Researchers have devised various scoring functions; however, no one has yet derived a truly accurate, broadly applicable method. Scoring functions used in virtual screening often work well for some targets but not as well for others. To overcome this hurdle, some programs allow more than one function to be employed with a single program. Consensus scoring, which combines several scoring methods, has been found to be superior to the use of a single function in some cases [129].

Another strategy is to use a simple function to discriminate between alternative binding geometries for a single ligand and combine it with more elaborate calculations to predict binding affinities. "Physics-based" methods to estimate binding affinity are computationally expensive but more accurate [102, 130, 131], but here we are moving into the realm of theoretical chemistry as opposed to cheminformatics.

The many docking programs currently available are usually judged (apart from speed) in terms of pose accuracy and enrichment (the ratio of the observed fraction of active compounds in the top few percent of a virtual screen to that expected by random selection). A large number of comparative and validation studies have been carried out [128, 132, 133]; progress has been aided by the availability of useful data sets such as ZINC, which contains over 13 million purchasable compounds in ready-to-dock, 3D formats [111], and the Directory of Useful Decoys (DUD), derived from ZINC, for benchmarking virtual screening [134, 135]. DUD contains a total of 2,950 active compounds against a total of 40 targets, and for each active, 36 "decoys" with similar physical properties (e.g., molecular weight and calculated Log*P*) but dissimilar topology. Data from the World of Molecular

Bioactivity (WOMBAT) database have also been used in conjunction with DUD [136].

Carrying out a valid comparative study is fraught with difficulties [128, 132, 137, 138]. Factors include the versions of the programs, the settings employed, fine tuning of parameters, quality of the data sets, preparation of receptors and ligands, and the criterion used for measuring accuracy. Using the root mean-square deviation (RMSD) as a criterion has been questioned [139]. There is also debate about enrichment factors and ROC curves [140, 141]. Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) [142] and robust initial enhancement (RIE) [143] have been proposed as alternatives to tackle the "early recognition" problem.

Several authors have found docking methods to be no more effective, or indeed less effective, than ligand-based virtual screening methods based on 3Dshape matching or 2D descriptors [144–146]. Several direct comparisons of docking with the shape-based tool ROCS [147] have been conducted using data sets from some recent docking publications. The results show that a shape-based, ligand-centric approach is more consistent than, and often superior to, the protein-centric approach taken by docking [144].

One way to improve a program's performance is to provide additional information such as pharmacophore(s) to orient the docking [148]. Self-docking is a good indication of a program's ability to identify native poses amongst several others, but provides little information about the accuracy in a real drug discovery scenario. For medicinal chemists, the best indicator of a program's accuracy is its ability to identify novel compounds that are then experimentally confirmed [131]. Apart from the need for better scoring functions, challenges remaining in protein–ligand docking include balancing speed and accuracy in conformational sampling; protein flexibility; loss of entropy; and protein desolvation [128, 149].

8. De Novo Design

De novo design involves the design of inhibitors from scratch given the target binding site. Typical programs include LUDI [150], SPROUT [151, 152], and BOMB [102, 131, 153]. More than 20 programs were reported in the literature in the early 1990s, but there was limited uptake of such systems because they designed molecules which were not easy to synthesize. De novo design is now popular again, but there is more consideration of generating molecules that are synthetically accessible [154] and which also represent non-obvious structural transformations.

Scores based on molecular complexity and retrosynthetic analysis are used to prioritize structures during generation. In these systems, molecular transformations are driven by known reactions. Synthetic feasibility is implicit in the rules, often based on a limited number of hand-picked reactions (typically derived from retrosynthetic analysis), and atom mapping is required. Gasteiger's team has devised a scoring method that rapidly evaluates synthetic accessibility of structures based on structural complexity, similarity to available starting materials, and assessment of strategic bonds where a structure can be decomposed to obtain simpler fragments. These individual components are combined to give an overall score of synthetic accessibility by an additive scheme [152, 155, 156]. Gillet and co-workers have developed a knowledge-based approach to de novo design which is based on reaction vectors that describe the structural changes that take place at the reaction center, along with the environment in which the reaction occurs [157].

9. Fragment-Based Drug Design

In fragment-based drug design (FBDD), molecules having a molecular weight of approximately 110–300 Da (smaller than the usual drug molecule) are used in structure-based drug design. Interest in this approach began in the mid-1990s when workers at Abbott, using "SAR by NMR," proved that meaningful SAR and stable binding modes could be observed even with mM ligands [158]. At around that same time, X-ray crystallography was beginning to be used to map out "hot spots" in protein binding sites [159]. Thus, FBDD was born [160–162]. Both NMR and X-ray analyses provide structural information about the binding site of a hit. A concurrent theme is the pursuit of leadlikeness rather than druglikeness [79–81] discussed in Subsection 2.

FBDD has two main advantages. Firstly, chemical space can be more efficiently probed by screening collections of small fragments rather than libraries of larger molecules: HTS of a million compounds covers only a tiny proportion of the available chemical space of 10^{60} or more compounds, whereas virtual screening of 10,000 fragments covers a much higher *proportion* of chemical diversity space. Less complex molecules should also show higher hit rates [81, 163]. The second major concept of FBDD concerns binding efficiency. The ligand efficiency (LE) measure, binding energy divided by the number of heavy atoms [164], and other measures [165–167] have been devised. The "Astex Rule of 3" (three or fewer hydrogen bond acceptors, three or fewer hydrogen bond donors, and CLog$P \leq 3$) has been suggested for selecting suitable fragments [168]. Most corporate collections contain

molecules that have been optimized for historical targets. HTS sometimes fails to find hits that interfere with new targets, such as protein–protein interaction surfaces. FBDD is robust for novel and difficult target classes although the success rate is likely to be dramatically lower than for easier targets. FBDD can also find new binding modes and allosteric sites.

Once hits have been found they must be fully characterized before crystallography or SAR studies are carried out; there may be many false positives. Surface plasmon resonance (SPR) and NMR can be used to eliminate non-specific binders. It is usually assumed that determining the X-ray structure of the fragment and target is an essential next step, but NMR and modeling may be used if an X-ray structure cannot be found. Structure-based drug design can then be used in optimization of the fragment [169].

NMR is still the commonest way of finding leads. Unfortunately, both NMR and X-ray crystallography screening require high concentrations of both fragments and target: the fragments must be soluble at high concentrations. SPR has the advantage of providing quantitative dynamics data on the binding interaction, such as binding constants, which are complementary to the structural information from X-ray and NMR screens. Some companies have also used high concentration bioassays, thermal methods, mass spectrometry (MS), and MS plus tethering with extenders (small molecules that bind to an active-site cysteine and contain a free thiol) [170]. Orthogonal validation (using two or even three assay methods in parallel) is a fairly new trend.

It has been claimed that a fragment screen provides a rapid and reliable means of interrogating a protein target for drugability before investing in further discovery research [169]. Drugability score (DScore) calculated by Schrödinger's SiteMap has allowed targets of low and high hit rates to be differentiated [163].

There have been a number of reports of the use of high-concentration screening or "reduced complexity" screening on compound collections that are a hybrid of a true fragment library and of a typical HTS collection. The upper molecular weight limit, for example, may be up to 350 Da. The available subset of molecules that can be screened within a corporate collection is thus increased, and these larger molecules, will, if active, be detectable in a HTS campaign simply by screening at a higher than normal concentration. On the other hand it has been argued that with this technique, a much larger library of leadlike compounds will be required to achieve a hit rate comparable to that observed for small fragments screened using very sensitive techniques at a higher concentration, and that at lower concentrations the smallest fragments will only be detectable if they have potency similar to that of the larger, more complex compounds being screened [167]. Publications describing the design and characterization of fragment libraries are becoming more common [167, 171].

One problem is that fragment-based approaches identify and characterize only "hot spots," i.e. the regions of a protein surface that are major contributors to the ligand binding free energy. Unfortunately, many binding sites in the active site that are responsible for target specificity and/or selectivity are not included in these "hot spots." Fragment screening finds the most efficient binder in the smallest core but optimization is still needed.

Once a hit has been found, it is optimized into a lead by one of three approaches: linking, growing, or merging. Linking may appear to be a good approach (energetically) but it can be hard to find second site fragments and affinity can be lost in conformational strain of the linker. It is usually more successful to grow fragments by structure-guided medicinal chemistry or grow by using the fragment binding motif to search for similar compounds that can be purchased (a method often called "SAR by catalog"). The merging approach has not been widely adopted. FBDD does not make drug design easier but it does offer more options. Optimization is still difficult and in some cases it is an intractable problem. Research continues on in situ fragment assembly techniques such as click chemistry, dynamic combinatorial library design, and tethering with extenders. A number of "success stories" have been published [169, 172], but as yet there is no "FBDD drug" on the market.

10. Molecular Similarity Analysis and Maximal Common Substructure

A very recent review covers a range of similarity methods in cheminformatics [173]; novel approaches to molecular similarity analysis have also been reviewed recently [174]. Some of the novel methods relate to QSAR and are discussed in the next section. Willett has reviewed advances in similarity based screening using 2D fingerprints [84]. Many different similarity coefficients have been used, including Tanimoto, cosine, Hamming, Russell-Rao, and Forbes. Willett's team has studied the use of data fusion methods: the similarity fusion approach is akin to consensus scoring in docking. In group fusion, not just one reference structure, but several structurally diverse reference structures are used [84]. The team has also worked on so-called turbo similarity searching which seeks to increase the power of the search engine by using a single reference structure's nearest neighbors.

Arguably, the most obvious way to compute the similarity between two 2D structures is to compare their maximal common substructures (MCS). An MCS is a single contiguously connected common subgraph of a molecular structure present in a specific

fraction of all molecules. Unfortunately, MCS isomorphism algorithms are extremely time-consuming [173]. Algorithms for finding an MCS include clique detection and clustering [175]. MCS is a core component of commercially available packages [176] such as Accelrys' Pipeline Pilot [177], the Simulations Plus ClassPharmer program [178], and ChemAxon's Library MCS clustering [179].

11. Quantitative Structure–Activity Relationships

Classical 3D QSAR methods, such as CoMFA, often provide accurate prediction of biological activity. Moreover, CoMFA models are interpretable, suggesting chemical changes that will improve biological activity. However, these methods can require time-consuming, expert preparation (molecular alignment, and conformer selection). Topomer CoMFA [180–182] minimizes the preparation needed for 3D QSAR analysis through an objective and consistent set of alignment rules. Topomer CoMFA can be used in conjunction with Topomer Search to identify the substituents and R-groups that are predicted to optimize the activity of compounds.

QSAR is no longer modeled in just one, two, or three dimensions [183]. 4D QSAR was an early approach to solving the alignment problem in 3D QSAR [184]. In 1D-QSAR, affinity is correlated with physicochemical properties such as pK_a and $\log P$; in 2D-QSAR it is correlated with 2D chemical connectivity; in 3D-QSAR with the three-dimensional structure; in 4D-QSAR with multiple representations of ligand conformation and orientation (as well as 3D structure); in 5D-QSAR with multiple representations of induced-fit scenarios (as well as with 4D concepts); and in 6D-QSAR with multiple representations of solvation models (as well as with 5D terms).

It might be assumed that the main objective of a QSAR study is to predict, for example, whether an untested compound will be active or inactive but, in practice, much work has been devoted to "explanatory" QSAR, relating changes in molecular structure to changes in activity, and only recently has there been considerable interest in predictivity. There are many reasons why models fail [185, 186], not least bad data, bad methodology, inappropriate descriptors, and domain inapplicability [187]. Significant issues concerning accuracy of prediction are extrapolation (whether the model can be applied to molecules unlike those in the training set) and overfitting [188]. Running cross-validation studies on the data is a reasonable check for overfitting but it is inadequate as a measure of extrapolation [189].

The outcome of a leave-one-out (LOO), or leave-many-out, cross-validation procedure is cross-validated R^2 (LOO q^2). The inadequacy of q^2 as a measure of predictivity was realized more than 10 years ago, in what has been referred to as "the Kubinyi paradox": models that give the best retrospective fit give the worst prospective results [190]. The "best fit" models are not the best ones in external prediction because internal predictivity tries to fit compounds in the training set as well as possible and does not take new compounds into account [191]. Even in the absence of real outliers, external prediction will be worse than fit because the model tries to "fit the errors" and attempts to explain them [186].

While a high value of q^2 is a necessary condition for high predictive power, it is not a sufficient condition. Tropsha and coworkers have argued that a reliable model should be characterized by both high q^2 and a high correlation coefficient (R^2) between the predicted and observed activities of compounds from a test set [192, 193]. They have proposed several approaches to the division of experimental data sets into training and test sets and have formulated a set of general criteria for the evaluation of the predictive power of QSAR models. Other reasons for overestimating q^2 are redundancy in the training set, or, in the case of nonlinear methods, the existence of multiple minima [193].

Doweyko [194] concludes that predictions can be enhanced when the test set is bounded by the descriptor space represented in the training set. Gramatica has discussed principles to define the validity and applicability domain of QSAR models [195], and in particular, emphasizes the need for external validation using at least 20% of the data. Validation is essential for application and interpretation of QSAR models [187, 196] and this necessity has been accepted by leading journals [197–199].

Researchers at Merck [189] have proposed a way to estimate the reliability of the prediction for an arbitrary chemical structure, using a given QSAR model, given the training set from which the model was derived. They found two useful measures: the similarity of the molecule to be predicted to the nearest molecule in the training set and the number of neighbors in the training set, where neighbors are those more similar than a user-chosen cut-off. Nevertheless, incorrect predictions of activity still arise among *similar* molecules even in cases where overall predictivity is high, because in Maggiora's well known metaphor, activity landscapes are not always like gently rolling hills, but may be more like the rugged landscape of the Bryce Canyon [200]. Even very local, linear models cannot account satisfactorily for landscapes with lots of "cliffs," and perfectly valid data points located in cliff regions may *appear* to be outliers, even though they are perfectly valid data points.

Following Maggiora's observations, there has been research into "activity landscape" characterization. The success of ligand-based virtual screening is much influenced by the nature of target-specific

structure–activity relationships, making it hard to apply computational methods consistently to recognize diverse structures with similar activity [174]. The performance of similarity-based methods depends strongly on the compound class that is studied, and approaches of different design and complexity often produce, overall, equally good (or bad) results. Moreover, there is often little overlap in the similarity relationships detected by different approaches, so alternative similarity methods need to be developed. SARs for diverse sets of active compounds or analog series are determined by the underlying "activity landscapes" [201].

On the basis of systematic correlation of 2D structural similarity and compound potency, Peltason and Bajorath have developed "a structure activity index (SARI)" that quantitatively describes the nature of SARs and establishes different SAR categories: continuous, discontinuous, heterogeneous-relaxed, and heterogeneous-constrained. Given a set of active compounds and their potency values, SAR Index calculations can estimate the likelihood of identifying structurally distinct molecules having similar activity [202].

Guha and van Drie have also studied activity cliffs [203, 204]. By use of a quantitative index, the structure–activity landscape index (SALI), they have identified pairs of molecules which are most similar but have the largest change in potency and hence form activity cliffs. They have shown how this provides a graphical representation of the entire SAR (where each node is a molecule, and each edge represents an activity cliff of varying magnitude), allowing the salient features of the SAR to be quickly grasped. Consensus activity cliffs [205] have also been described.

It has been said that a general feeling of disillusionment with QSAR has settled across the modeling community [206] but actually there is renewed interest in QSAR in the field of absorption, distribution, excretion, metabolism, and toxicity (ADMET); many regulatory laws including the new Registration, Evaluation, Authorization of Chemicals (REACH) legislation in Europe have prompted significant new activity [207–209]. Under REACH regulation, information on intrinsic properties of substances may be generated by means other than tests, provided that certain conditions are met, so animal testing can be reduced or avoided by replacing traditional test data with predictions or equivalent data. Integrated testing strategies, including in vitro assays, QSARs, and "read-across," can be used in a combined "non-testing" strategy, i.e. as an alternative to the use of animals. In read across, known information on the property of a substance is used to make a prediction of the same property for another substance that is considered similar. This avoids the need to test every substance for every endpoint, but there are conditions. QSARs are allowed under REACH if the method is scientifically

valid, the domain is applicable, the endpoint is relevant, and adequate documentation is provided [207, 208].

Consensus QSAR has been applied to models developed from different techniques and different data sets; the method often performs better than a single QSAR [210]. One study by Cronin's team, however, shows that the use of consensus models does not seem warranted given the minimal improvement in model statistics for the data sets in question [211]. The Food and Drug Administration (FDA) has used multiple commercially available programs, with the same data sets, to predict carcinogenicity [212]. The FDA has several reasons for using more than one QSAR software program. None of the programs has all the necessary functionalities, and none has 100% coverage, sensitivity, and specificity. All of the programs are complementary. The individual models made complementary predictions of carcinogenesis and had equivalent predictive performance. Consensus predictions for two programs achieved better performance, better confidence predictions, and better sensitivity. Consensus models from three different expert systems have also been used with some success in prediction of mutagenicity using the commercial system KnowItAll [213]. Consensus models have been tailored to a risk assessment scenario in AstraZeneca [214]. There are, however, disadvantages. Consensus models hide outliers, incorrect data, and interesting parts of the data set. They lack portability, transparency, and mechanistic interpretation.

12. Metabolite Prediction

Various rule-based and statistical methods have been used to predict metabolic fate. Gasteiger's team [215] has used descriptors of drugs metabolized by human cytochrome P450 (CYP) isoforms 3A4, 2D6, and 2C9 in model building methods such as multinomal logistic regression, decision tree, or support vector machine (SVM). This team also supplies the biochemical pathways database, BioPath [216]. Schwaighofer and co-workers [217] have developed machine learning tools to predict the metabolic stability of compounds from drug discovery projects at Bayer Schering. They concluded that Gaussian Process classification has specific benefits.

Another team [214] has used data mining methods to exploit biotransformation data that have been recorded in the Symyx Metabolite [218] database. Reacting center fingerprints were derived from a comparison of substrates and their corresponding products listed in the database. The metabolic reaction data were then mined by submitting a new molecule and searching for

fingerprint matches. An "occurrence ratio" was derived from the fingerprint matches between the submitted compound and the reacting center and substrate fingerprint databases. The method enables the results of the search to be rank-ordered as a measure of the relative frequency of a reaction occurring at a specific site within the submitted molecule.

The rule-based method, Systematic Generation of Metabolites (SyGMa) [219] predicts potential metabolites based on reaction rules derived from metabolic reactions that occur in man, reported in Symyx' Metabolite database [218]. An empirical probability score is assigned to each rule representing the fraction of correctly predicted metabolites in the training database. This score is used to refine the rules and to rank predicted metabolites. Another team [220] has used absolute and relative reasoning to prioritize biotransformations, since they argue that a system which predicts the metabolic fate of a chemical should predict the more likely metabolites rather than every possibility.

13. Pharmacophores

Three-dimensional pharmacophore methods in drug discovery have been reviewed very recently [221]. The technologies used in 3D pharmacophore modeling packages such as Accelrys' Catalyst, Chemical Computing Group's Molecular Operating Environment (MOE), Schrödinger's Phase, and Inte:ligand's LigandScout have also been reviewed recently [222]. Another method, PharmID, uses fingerprints of 3D features and a modification of Gibbs sampling to align a set of known flexible ligands, where all compounds are active [223]. A clique detection method is used to map the features back onto the binding conformations. The algorithm is able to handle multiple binding mode problems, which means it can superimpose molecules within the same data set according to two different sets of binding features.

Alignment of multiple ligands is particularly difficult when the spatial overlap between structures is incomplete, in which case no good template molecule is likely to exist. Pairwise rigid ligand alignment based on linear assignment (the LAMDA algorithm) has the potential to address this problem [224]. A version of LAMDA is embodied in the program named Genetic Algorithm with Linear Assignment for Hypermolecule Alignment of Datasets (GALAHAD) developed by Tripos in collaboration with Biovitrum and Sheffield University [225]. GALAHAD creates pharmacophore and alignment models from diverse sets of flexible, active compounds, generating models that are sterically, pharmacophorically, and energetically optimal. It supports partial

matching of features and partial coverage. By decoupling conformational searching from alignment the program frees scientists from the need to fit all ligands to any single template molecule. GALAHAD uses a multi-objective optimization (Pareto ranking) scoring function. A single Pareto run produces better, more diverse models than an entire series of GA runs using a range of fitness term weights [78]. Tripos also offers the Surflex-Sim [226] method of molecular alignment and virtual screening.

Other new programs have become commercially available. Chemical Computing Group's MOE contains a pharmacophore elucidator that takes a collection of actives and inactive analogs, generates all pharmacophores (there may be multiple correct answers to a pharmacophore elucidation procedure) and validates each to see which one is best. Schrödinger's Phase is a package of pharmacophore modeling tools that offers scientists control at each step (including pharmacophore scoring, QSAR building, and database screening) and enables users to modify existing feature definitions and create new features. Some programs may be able to uncover multiple binding modes. The ability to use larger data sets, the coverage of conformational space, and the ability to handle more flexible molecules are reported to be other advantageous features.

High quality data sets have driven progress in the field of protein ligand docking but there are fewer data sets for pharmacophore mapping. It has been suggested that too many algorithms could have been developed using the same data, hampering progress in the field [227]. The Patel data set [228] has been used to evaluate Catalyst [53, 54], DISCO [52] and GASP [55–57] and has been refined [225] in the development of GALAHAD and in the Multi-Objective Genetic Algorithm (MOGA) program [77, 78]. MOGA looks for solutions which are compromises between three objective functions: energy, volume overlap, and feature score. It is now being tested on a new data set, the "Taylor" data set, which aims to cover 25–35 proteins, and is carefully compiled by an expert in the field. Protonation states are checked, and ligand geometries and electron density in the binding site are being checked. A detailed analysis of crystal structures is carried out to elucidate pharmacophore points. Each of the 10 protein targets has a minimum of two ligands and a maximum of 16 ligands [227].

Clark has discussed synergy between ligand-based and structure-based methods [229]. Unfortunately, ligand binding often induces structural changes that significantly reduce the usefulness of apoprotein structures for docking and scoring. In such cases it is often better to dock into the binding site of a ligand–protein complex from which the ligand has been extracted *in silico*. Even when a native protein structure is suitable for docking, ligands can

provide critical information about the location of the relevant binding site. Moreover, interactions with specific binding site residues illuminated by bound ligands have been successfully used to direct docking and to tailor scoring functions to specific target proteins. An extreme version of this is the use of docking to align molecules for CoMFA. Target-based methods such as FlexX are moving in a ligand-based direction, and CoMFA and CoMSIA (ligand-based methods) are moving towards a target-based approach.

The two approaches have also been combined by parallel application and in series; the former is a form of consensus scoring whereas the latter has been called "consensus screening." Scoring is used with surface feature complementarity in eHiTS [89]. Generation of a Structural Interaction Fingerprint (SIFt) [230] translates 3D structural binding information from a protein–ligand complex into a one-dimensional binary string. Each fingerprint represents the "structural interaction profile" of the complex that can be used to organize, analyze, and visualize the rich amount of information encoded in ligand–receptor complexes and also to assist database mining. Muegge and Oloff [231] have also discussed synergies between structure-based and ligand-based virtual screening.

14. Data Reduction and Visualization

Computer-aided drug design produces huge volumes of data, which are often multidimensional in nature. There is, thus, a need for methods to reduce the dimensionality of these data and visualize the results [232, 233]. Visualization techniques include self-organizing maps [44, 234, 235], tree maps [236–238], enhanced SAR Maps [44], dendrograms [239], radial clustergrams [240], non-linear maps [241–243], heatmaps [237], and various forms of conventional statistical plots (scatter plots, bar charts, pie charts, etc.). SAR trees [244] and scaffold trees [245, 246] make use of common substructures.

A SAR tree represents a collection of compounds as an acyclic graph, where each node represents a common substructure and its children represent the R-groups around it. Each R-group in turn embodies another common substructure that is shared by multiple compounds at that particular attachment site, and is recursively split into more refined R-groups, until there are no further variations. This method is also used in the commercially available ClassPharmer software [178].

Rule-based methods such as that of Bemis and Murcko [247] scale linearly with the number of structures since the classification

process is done individually for each molecule and incremental update is possible. The classes created by such methods are more intuitive to chemists than those produced by clustering and other methods. Chemical Abstracts Service has expanded on Bemis and Murcko's frameworks technique in developing SubScape, a new substance analysis and visualization tool for substances retrieved in SciFinder [248]. The SubScape Framework Identifier combines three identification numbers, each of which signifies one aspect of framework structure: a graph id denotes the underlying connectivity, a node id denotes the pattern of elements, and a bond id denotes the pattern of bond types.

The scaffold tree technique reported by Schuffenhauer and co-workers is also a variation on Bemis and Murcko's molecular frameworks. This hierarchical classification method [246] uses molecular frameworks as the leaf nodes of a scaffold tree. By iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first. Highlighting by color intensity is used to show the fraction of active compounds containing a scaffold: this immediately identifies those branches of the scaffold tree which contain active molecules. Schuffenhauer and his colleagues [249] have compared rule-based and scaffold-oriented classification methods (in a Pareto analysis), and clustering based on molecular descriptors: no technique was found to be generally superior but all gave results which were to some extent biologically meaningful.

Agrafiotis et al. [250] have developed SAR maps to allow medicinal chemists to visualize structure-activity relationships. An SAR map renders an R-group decomposition of a congeneric series as a rectangular matrix of cells, each representing a unique combination of R-groups, color-coded by a property of the corresponding compound. An enhanced version of the software [251] expands the types of visualizations that can be displayed inside the cells. Examples include multidimensional histograms and pie charts that visualize the biological profiles of compounds across a panel of assays, forms that display specific fields on user-defined layouts, aligned 3D structure drawings that show the relative orientation of different substituents, dose–response curves, and images of crystals or diffraction patterns.

Medicinal chemists at ArQule can use the Spiral View tool [252] developed by Smellie. The relationship depiction paradigm of this tool is a spiral view centered on the most active compound, with the compounds most similar to it oriented clockwise around it. The chemist "walks around" the spiral. The width of the line between two compounds is proportional to the difference in property values between them. When the user clicks on a molecule, it then becomes the central molecule in a new spiral.

15. Text Mining

The discussion so far has covered many data mining techniques but drug-related information is also buried within written resources. Text mining is a relatively new technique used to extract this information automatically [253–258]. The first stage in this is information retrieval (IR) from the scientific literature, or patents, or in-house documents, which is carried out with general search engines or more specialized IR tools. The next step is information extraction (IE) from the text retrieved. Approaches to IE include rule-based or knowledge-based methods and statistical or machine-learning based methods. Named entity extraction (NER) recognizes terms for genes, proteins, drugs etc. NER can be based on dictionaries, rules, training sets (in machine learning) and combinations of these approaches. NER may involve simpler techniques such as co-occurrence of terms in the text, or the more sophisticated techniques of natural language processing.

In a next step, these extracted chemical names are converted into connection tables by means of a commercially available program (ACD/Labs, ChemAxon, CambridgeSoft, OpenEye Scientific Software, and InfoChem supply such software) and may be stored and retrieved in a chemical structure database system. A number of commercial organizations (e.g., TEMIS, Linguamatics, Notiora, IBM, SureChem, and InfoChem) have developed algorithms for chemical named entity recognition and have established relationships with cheminformatics companies such as Symyx, Accelrys, ChemAxon, and InfoChem. The publishing group of the Royal Society of Chemistry (RSC) has used the NER software Open Source Chemical Analysis Routines (OSCAR, developed at the University of Cambridge) to enhance articles with "live" chemical structures, in its system RSC Prospect [259–262]. RSC has used selections from the Open Biomedical Ontologies and it makes its own chemical ontologies freely available [263].

A more complex problem is the analysis of structure images, or "chemical optical character recognition": recognizing and distinguishing different graphical objects in a picture (structures, arrows, and text), and the conversion of the structure drawings into connection tables. To this end, four programs are currently under active development: Chemical Literature Date Extraction, CLiDE [264], chemoCR [265], Optical Structure Recognition Software, OSRA [266], and ChemReader [267]. Analyzing the image of a Markush structure is a challenge for the future: in this case there is the additional problem of finding the appropriate R-groups from the part of the text of the article or patent which acts as a key to the diagram in question.

16. Conclusion

Research continues into other applications involving Markush structures. For example, cheminformatics programs could be developed to include "freedom to operate" patent space in optimization studies. Text mining is still in its infancy: more chemical dictionaries and ontologies are constantly being developed and the performance of image recognition programs is gradually improving. At least three teams, at Key Module [268], Molecular Networks [39], and InfoChem [269] are active in the field of retrosynthesis. In computer-aided drug design, there is likely to be keen interest in machine learning in future: random forest seems to be gaining in popularity.

Matthews [270] has suggested many unmet needs in QSARs and expert systems, including integrated fragment and descriptor paradigms and 3D descriptors; QSARs based upon pure active ingredient and metabolites; QSARs for drug–drug interaction, for animal organ toxicities, and for regulatory dose concentration endpoints (e.g. lowest observed effect level and no observed effect level); and expert system rules for toxicities of substances such as biologicals which cannot be predicted by QSAR. Other unmet needs are databases of pharmaceutical off-target activities of pharmaceutical investigational new drugs, of confidential business information, and of regulatory dose concentration endpoints; integration of FDA and Environmental Protection Agency archival data; and advanced linguistic software to extract data.

Some of the challenges of structure-based drug design (docking and de novo design) have already been discussed. Prediction of affinity is a very hard problem and is as yet unsolved; Jorgensen believes that free energy guided molecular design, for example, may become a mainstream activity [131]. Ligands forming covalent complexes have been little studied. Proteins have been the major targets of docking methods. However, nucleic acids are also targets for medicinal chemistry and should be further investigated. Docking DNA intercalators is even more challenging [128].

In fragment-based drug design there are still many opportunities for new development, such as improving the novelty, structural diversity and physicochemical properties of fragment libraries, and improving detection methods to find hits with activity as low as 1–10 mM. It should be possible to identify new types of interactions: protein–protein interactions, novel templates, and new binding modes. There is much work to be done on increasing the efficiency of fragment optimization. One challenge is deciding which fragments to progress, other than using the subjective decisions of a medicinal chemist. Tools for assessing synthetic

accessibility may help. Optimizing fragments in the absence of a crystal structure is another hurdle. Progress in structural biology will lead to progress in FBDD.

Target-based and ligand-based methods have shared challenges. Targets are not rigid: account must be taken of alternative binding sites, tautomeric ambiguity, and accommodation; local and global binding site plasticity; and complex librational freedom. Ligands move too: the problem of discrete or complete conformational sampling is solved but tautomerism and pK_a are unsolved problems. Solvation and (de)solvation of ligands and targets is another shared challenge. Above all there is the enormous problem of entropy [271].

Practitioners of bioinformatics and cheminformatics have much to learn from each other. Oprea and colleagues have coined the term "systems chemical biology," believing that the future of cheminformatics lies in its ability to provide an integrative, predictive framework that links biological sciences [104]. Others are publishing in that field [272]. Computational chemical biology [273] certainly has a bright future.

References

1. Warr, W. A. Cheminformatics education. http://www.qsarworld.com/cheminformatics-education.php (accessed October 2, 2009).
2. Willett, P. (2008) A bibliometric analysis of the literature of chemoinformatics. *Aslib. Proc.* **60**, 4–17.
3. Cheminformatics or chemoinformatics? http://www.molinspiration.com/chemoinformatics.html (accessed October 2, 2009).
4. Warr, W. A. (1999) Balancing the needs of the recruiters and the aims of the educators, in *Book of Abstracts, 218th ACS National Meeting, New Orleans, Aug. 22–26.*
5. Warr, W. A. Extract from 218th ACS National Meeting and Exposition, New Orleans, Louisiana, August 1999 [cheminformatics]. http://www.warr.com/warrzone2000.html (accessed October 2, 2009).
6. Willett, P. (2007) A bibliometric analysis of the Journal of Molecular Graphics and Modelling. *J. Mol. Graphics Modell.* **26**, 602–606.
7. Leach, A. R., and Gillet, V. J. (2003) *An Introduction to Chemoinformatics.* Kluwer, Dordrecht, The Netherlands.
8. Gasteiger, J., and Engel, T., (Eds.) (2003) *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, Germany.
9. Bajorath, J., (Ed.) (2004) *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, Humana Press, Totowa, NJ.
10. Oprea, T. I., (Ed.) (2005) *Chemoinformatics in Drug Discovery*, Wiley, New York, NY.
11. Schneider, G., and Baringhaus, K.-H., (Eds.) (2008) *Molecular Design: Concepts and Applications*, Wiley-VCH, Weinheim, Germany.
12. Chen, W. L. (2006) Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.* **46**, 2230–2255.
13. Engel, T. (2006) Basic overview of chemoinformatics. *J. Chem. Inf. Model.* **46**, 2267–2277.
14. Willett, P. (2008) From chemical documentation to chemoinformatics: 50 years of chemical information science. *J. Inf. Sci.* **34**, 477–499.
15. Willett, P. (2003) A history of chemoinformatics in *Handbook of Chemoinformatics: From Data to Knowledge* Wiley-VCH, Weinheim, Germany, Vol. 1, pp 6–20.
16. Bishop, N., Gillet, V. J., Holliday, J. D., and Willett, P. (2003) Chemoinformatics research at the University of Sheffield: a history and citation analysis. *J. Inf. Sci.* **29**, 249–267.
17. Ash, J. E., Warr, W. A., and Willett, P., (Eds.) (1991) *Chemical structure systems: computational techniques for representation, searching, and processing of structural information*, Ellis Horwood, Chichester, UK.

18. Paris, G. C. (1997) Chemical structure handling by computer. *Ann. Rev. Inf. Sci. Technol.* **32**, 271–337.
19. Paris, G. C. (1998) Structure databases, in *Encyclopedia of Computational Chemistry* (Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, and Schreiner, P. R., Eds.), Wiley, Chichester, UK, Vol. 4, pp 2771–2785.
20. Paris, G. C. (2003) Databases of chemical structures, in *Handbook of Cheminformatics: From Data to Knowledge* (Gasteiger, J., Ed.), Wiley-VCH, Weinheim, Germany, Vol. 2, pp 523–555.
21. Warr, W. A., and Suhr, C. (1992) *Chemical information management*, Wiley-VCH, Weinheim, Germany.
22. Ott, M., A. (2004) Cheminformatics and organic chemistry. Computer-assisted synthetic analysis, in *Cheminformatics Developments* (Noordik, J., H., Ed.), IOS Press, Amsterdam, The Netherlands, pp 83–109.
23. Pearlman, R. S. (1987) Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Autom. News* **2**, 5–7.
24. Gasteiger, J., Rudolph, C., and Sadowski, J. (1990) Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **3**, 537–547.
25. Hiller, C., and Gasteiger, J. (1987) An automatic molecule builder, in *Software Development in Chemistry 1. Proceedings of the Workshops on the Computer in Chemistry, Hochfilzen/Tirol, November 19–21, 1986* (Gasteiger, J., Ed.), Springer, Berlin, Vol. 1, pp 53–66.
26. Sadowski, J., Gasteiger, J., and Klebe, G. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008.
27. Jakes, S. E., and Willett, P. (1986) Pharmacophoric pattern matching in files of 3-D chemical structures: selection of interatomic distance screens. *J. Mol. Graphics* **4**, 12–20.
28. Jakes, S. E., Watts, N., Willett, P., Bawden, D., and Fisher, J. D. (1987) Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance. *J. Mol. Graphics* **5**, 41–48.
29. Brint, A. T., and Willett, P. (1987) Pharmacophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics* **5**, 49–56.
30. Cringean, J. K., Pepperrell, C. A., Poirrette, A. R., and Willett, P. (1990) Selection of screens for three-dimensional substructure searching. *Tetrahedron Comput. Methodol.* **3**, 37–46.
31. Willett, P. (1991) *Three-dimensional Chemical Structure Handling*, Research Studies Press, Taunton, UK.
32. Martin, Y. C., and Willett, P., (Eds.) (1998) *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, American Chemical Society, Washington, DC.
33. Martin, Y. C., Danaher, E. B., May, C. S., and Weininger, D. (1988) MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J. Comput.-Aided Mol. Des.* **2**, 15–29.
34. Van Drie, J. H., Weininger, D., and Martin, Y. C. (1989) ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **3**, 225–251.
35. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
36. Barnard, J. M. (1991) A comparison of different approaches to Markush structure handling. *J. Chem. Inf. Comput. Sci.* **31**, 64–68.
37. Benichou, P., Klimczak, C., and Borne, P. (1997) Handling genericity in chemical structures using the Markush DARC software. *J. Chem. Inf. Comput. Sci.* **37**, 43–53.
38. Corey, E. J., and Wipke, W. T. (1969) Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192.
39. THERESA. Molecular Networks. http://www.molecular-networks.com/products/theresa (accessed October 2, 2009).
40. Dugundji, J., and Ugi, I. (1973) Algebraic model of constitutional chemistry as a basis for chemical computer programs. *Fortschr. Chem. Forsch.* **39**, 19–64.
41. Bauer, J., Herges, R., Fontain, E., and Ugi, I. (1985) IGOR and computer assisted innovation in chemistry. *Chimia* **39**, 43–53.
42. Gasteiger, J., Ihlenfeldt, W. D., Roese, P., and Wanke, R. (1990) Computer-assisted reaction prediction and synthesis design. *Anal. Chim. Acta* **235**, 65–75.
43. Jorgensen, W. L., Laird, E. R., Gushurst, A. J., Fleischer, J. M., Gothe, S. A., Helson, H. E., Paderes, G. D., and Sinclair, S. (1990) CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **62**, 1921–1932.
44. Sadowski, J., Wagener, M., and Gasteiger, J. (1996) Assessing similarity and diversity

of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem. Int. Ed. Engl.* **34**, 2674–2677.
45. Sadowski, J., Gasteiger, J., and Klebe, G. (2002) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008.
46. Kennard, O., Watson, D. G., and Town, W. G. (1972) Cambridge Crystallographic Data Centre. I. Bibliographic file. *J. Chem. Doc.* **12**, 14–19.
47. Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G., and Watson, D. G. (1973) Cambridge Crystallographic Data Centre. II. Structural data file. *J. Chem. Doc.* **13**, 119–123.
48. Allen, F. H. (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. Sect. B Struct. Sci.* **B58**, 380–388.
49. Allen, F. H., Battle, G., and Robertson, S. (2007) The Cambridge Structural Database, in *Comprehensive Medicinal Chemistry II.* (Triggle, D. J., and Taylor, J. B., Eds.), Elsevier, Amsterdam, The Netherlands, Vol. 3, pp 389–410.
50. Berman, H. M. (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr., Sect. A Found. Crystallogr.* **A64**, 88–95.
51. Martin, Y. C. (1998) Pharmacophore mapping, in *Designing Bioactive Molecules* (Martin, Y. C., and Willett, P., Eds.), American Chemical Society, Washington, DC, pp 121–148.
52. Martin, Y. C., Bures, M. G., Danaher, E. A., DeLazzer, J., Lico, I., and Pavlik, P. A. (1993) A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **7**, 83–102.
53. Greene, J., Kahn, S., Savoj, H., Sprague, P., and Teig, S. (1994) Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.* **34**, 1297–1308.
54. Sprague, P. (1995) Automated chemical hypothesis generation and database searching with Catalyst. *Perspect. Drug Discov. Des.* **3**, 1–20.
55. Jones, G., Willett, P., and Glen, R. C. (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **9**, 532–549.
56. Jones, G., Willett, P., and Glen, R. C. (1995) Molecular recognition of a receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53.
57. Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748.
58. Winkler, D. A. (2002) The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. *Briefings Bioinf.* **3**, 73–86.
59. Tropsha, A. (2003) Recent trends in quantitative structure-activity relationships, in *Burger's Medicinal Chemistry and Drug Discovery, Volume 1, Drug Discovery* (Abraham, D. J., Ed.) 6th ed., Wiley, New York, NY.
60. Tropsha, A. (2005) Application of predictive QSAR models to database mining, in *Cheminformatics in Drug Discovery* (Oprea, T. I., Ed.), Wiley, New York, NY, pp 437–455.
61. Gramatica, P. A short history of QSAR evolution. http://www.qsarworld.com/Temp_Fileupload/Shorthistoryofqsar.pdf (accessed September 23, 2009).
62. Hawkins, D. M. QSAR approaches, models and statistics relating to toxicity prediction. In W. A Warr. Proceedings of New Horizons in Toxicity Prediction. Lhasa Limited symposium event in collaboration with the University of Cambridge, December 2008. http://www.qsarworld.com/files/Lhasa_Symposium_2008_Report.pdf (accessed September 23, 2009).
63. Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959–5967.
64. Klebe, G., Abraham, U., and Mietzner, T. (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130–4146.
65. Lemmen, C., Lengauer, T., and Klebe, G. (1998) FlexS: a method for fast flexible ligand superposition. *J. Med. Chem.* **41**, 4502–4520.
66. Kennedy, T. (1997) Managing the drug discovery/development interface. *Drug Discov. Today* **2**, 436–444.
67. Benigni, R., (Ed.) (2003) *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, CRC Press, Boca Raton, FL.
68. Cronin, M. T. D., and Livingstone, D. J., (Eds.) (2004) *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton, FL.
69. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery

and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25.
70. Cramer, R. (1995) Unpublished work.
71. Güner, O. F., (Ed.) (2000) *Pharmacophore: Perception, Development, and Use in Drug Design. [In: IUL Biotechnol. Ser., 2000; 2]*, International University Line, La Jolla, CA.
72. Willett, P., (Ed.) (1997) *Computational Methods for the Analysis of Molecular Diversity (Perspectives in Drug Discovery and Design 1997, Volumes 7/8)*, Kluwer/Escom, Dordrecht, The Netherlands.
73. Brown, R. D., and Martin, Y. C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **36**, 572–584.
74. Brown, R. D., and Martin, Y. C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **37**, 1–9.
75. Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., and Moos, W. H. (1995) Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **38**, 1431–1436.
76. Gillet, V. J., Willett, P., and Bradshaw, J. (1997) The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **37**, 731–740.
77. Cottrell, S. J., Gillet, V. J., Taylor, R., and Wilton, D. J. (2004) Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *J. Comput.-Aided Mol. Des.* **18**, 665–682.
78. Cottrell, S., Gillet, V., and Taylor, R. (2006) Incorporating partial matches within multiobjective pharmacophore identification. *J. Comput.-Aided Mol. Des.* **20**, 735–749.
79. Teague, S., J., Davis, A., M., Leeson, P., D., and Oprea, T. I. (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* **38**, 3743–3748.
80. Oprea, T. I., Davis, A. M., Teague, S. J., and Leeson, P. D. (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **41**, 1308–1315.
81. Hann, M. M., Leach, A. R., and Harper, G. (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864.
82. Sadowski, J., and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**, 3325–3329.
83. Wagener, M., and Van Geerestein, V. J. (2000) Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* **40**, 280–292.
84. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053.
85. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288.
86. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**, 470–489.
87. McMeeking, B., and Fletcher, D. (2004) The United Kingdom Chemical Database Service: CDS, in *Cheminformatics Developments* (Noordik, J., H., Ed.), IOS Press, Amsterdam, The Netherlands, pp 37–67.
88. (Multiple authors) (2006) Focus on Cyberinfrastructure ("e-science"), the Enabling Platform for Cheminformatics. *J. Chem. Inf. Model.* **46**.
89. Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., and Johnson, A. P. (2007) eHiTS: a new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **26**, 198–212.
90. Salamone, S. GTA4: enabler of life sciences research? http://www.bio-itworld.com/inside-it/2008/2005/gta2004-and-life-sciences.html (accessed October 2, 2009).
91. Murray-Rust, P. (2008) Chemistry for everyone. *Nature (London, U. K.)* **451**, 648–651.
92. Williams, A. J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* **13**, 502–506.
93. Warr, W. A. (2008) Social software: fun and games, or business tools? *J. Inf. Sci.* **34**, 591–604.
94. Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S., and Zhang, Y. (2005) Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* **3**, 1832–1834.
95. Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J., and Willighagen, E. L. (2006) The Blue Obelisk. Interoperability in chemical informatics. *J. Chem. Inf. Model.* **46**, 991–998.
96. Taylor, K. R., Gledhill, R. J., Essex, J. W., Frey, J. G., Harris, S. W., and De Roure, D. C.

(2006) Bringing chemical data onto the semantic web. *J. Chem. Inf. Model.* **46**, 939–952.
97. Frey, J. G. (2009) The value of the Semantic Web in the laboratory. *Drug Discov. Today* **14**, 552–561.
98. Gardner, B. Approaches to information integration. Paper given at ICIC 2007, Sitges, Spain, October 21–24, 2007. http://www.infonortics.eu/chemical/ch07/slides/gardner.pdf (accessed September 18, 2009).
99. Symyx Technologies. DiscoveryGate. http://www.discoverygate.com (accessed October 2, 2009).
100. ChemSpider. http://www.chemspider.com (accessed September 18, 2009).
101. Warr, W. A. Workflow and pipelining in cheminformatics. http://www.qsarworld.com/qsar-workflow1.php (accessed October 2, 2009).
102. Jorgensen, W. L. (2004) The many roles of computation in drug discovery. *Science* **303**, 1813–1818.
103. PubChem. http://pubchem.ncbi.nlm.nih.gov/search/search.cgi (accessed September 18, 2009).
104. Oprea, T. I., Tropsha, A., Faulon, J.-L., and Rintoul, M. D. (2007) Systems chemical biology. *Nat. Chem. Biol.* **3**, 447–450.
105. Therapeutics for Rare and Neglected Diseases. http://www.nih.gov/news/health/may2009/nhgri-2020.htm (accessed October 7, 2009).
106. DeLano, W. L. (2005) The case for open-source software in drug discovery. *Drug Discov. Today* **10**, 213–217.
107. Stahl, M. T. (2005) Open-source software: not quite endsville. *Drug Discov. Today* **10**, 219–222.
108. The Pistoia Alliance. http://pistoiaalliance.org/ (accessed October 7, 2009).
109. Barnes, M. R., Harland, L., Foord, S. M., Hall, M. D., Dix, I., Thomas, S., Williams-Jones, B. I., and Brouwer, C. R. (2009) Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discov.* **8**, 701–708.
110. Lhasa Limited. Derek for Windows. http://www.lhasalimited.org/ (accessed October 7, 2009).
111. Irwin, J. J., and Shoichet, B. K. (2004) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182.
112. eMolecules. http://www.emolecules.com (accessed September 18, 2009).
113. NIH/CADD chemical structure lookup service. http://cactus.nci.nih.gov/lookup (accessed October 2, 2009).
114. Williams, A. J. (2008) A perspective of publicly accessible/open-access chemistry databases. *Drug Discov. Today* **13**, 495–501.
115. Fink, T., and Reymond, J.-L. (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **47**, 342–353.
116. Blum, L. C., and Reymond, J.-L. (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733.
117. Collaborative Drug Discovery. http://www.collaborativedrug.com/ (accessed September 18, 2009).
118. Morgan, H. L. (1965) The generation of a unique machine description for chemical structures – a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113.
119. Wipke, W. T., and Dyott, T. M. (1974) Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **96**, 4834–4842.
120. Hillard, R., and Taylor, K. T. (2009) InChI keys as standard global identifiers in chemistry web services, in *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, March 22–26, 2009.*
121. The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi/ (accessed September 18, 2009).
122. Heller, S. R., and McNaught, A. D. (2009) The IUPAC international chemical identifier (InChI). *Chem. Int.* **31**, 7–9.
123. Warr, W. A. The IUPAC International Chemical Identifier. http://www.qsarworld.com/INCHI1.php (accessed September 18, 2009).
124. McKay, B. D. (1981) Practical graph isomorphism. *Congressus Numeratium* **30**, 45–87.
125. Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
126. Weininger, D., Weininger, A., and Weininger, J. L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101.
127. Leach, A. R., Shoichet, B. K., and Peishoff, C. E. (2006) Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **49**, 5851–5855.

128. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., and Corbeil, C. R. (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153**, S7-S26.
129. Wang, R., Lai, L., and Wang, S. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **16**, 11–26.
130. Waszkowycz, B. (2008) Towards improving compound selection in structure-based virtual screening. *Drug Discov. Today* **13**, 219–226.
131. Jorgensen, W. L. (2009) Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**, 724–733.
132. Irwin, J. (2008) Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **22**, 193–199.
133. Liebeschuetz, J. (2008) Evaluating docking programs: keeping the playing field level. *J. Comput.-Aided Mol. Des.* **22**, 229–238.
134. DUD. A Directory of Useful Decoys. http://dud.docking.org/ (accessed September 18, 2009).
135. Huang, N., Shoichet, B. K., and Irwin, J. J. (2006) Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801.
136. Good, A., and Oprea, T. (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **22**, 169–178.
137. Jain, A. (2008) Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **22**, 201–212.
138. Jain, A., and Nicholls, A. (2008) Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **22**, 133–139.
139. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., and Taylor, R. (2005) Comparing protein-ligand docking programs is difficult. *Proteins Struct., Funct., Bioinf.* **60**, 325–332.
140. Clark, R., and Webster-Clark, D. (2008) Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **22**, 141–146.
141. Nicholls, A. (2008) What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **22**, 239–255.
142. Truchon, J.-F., and Bayly, C. I. (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **47**, 488–508.
143. Sheridan, R. P., Singh, S. B., Fluder, E. M., and Kearsley, S. K. (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* **41**, 1395–1406.
144. Hawkins, P. C. D., Skillman, A. G., and Nicholls, A. (2006) Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **50**, 74–82.
145. Zhang, Q., and Muegge, I. (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* **49**, 1536–1548.
146. McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.-F., and Cornell, W. D. (2007) Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **47**, 1504–1519.
147. Rush, T. S., Grant, J. A., Mosyak, L., and Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. *J. Med. Chem.* **48**, 1489–1495.
148. Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T. M., Murray, C. W., Taylor, R. D., and Watson, P. (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **44**, 793–806.
149. Klebe, G. (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **11**, 580–594.
150. Böhm, H.-J. (1992) The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **6**, 61–78.
151. Law, J. M. S., Fung, D. Y. K., Zsoldos, Z., Simon, A., Szabo, Z., Csizmadia, I. G., and Johnson, A. P. (2003) Validation of the SPROUT *de novo* design program. *Theochem* **666–667**, 651–657.
152. Boda, K., and Johnson, A. P. (2006) Molecular complexity analysis of *de novo* designed ligands. *J. Med. Chem.* **49**, 5869–5879.
153. Barreiro, G., Kim, J. T., Guimaraes, C. R. W., Bailey, C. M., Domaoal, R. A., Wang, L., Anderson, K. S., and Jorgensen, W. L. (2007) From docking false-positive to active anti-HIV agent. *J. Med. Chem.* **50**, 5324–5329.
154. Baber, J. C., and Feher, M. (2004) Predicting synthetic accessibility: application in drug discovery and development. *Mini-Rev. Med. Chem.* **4**, 681–692.
155. Boda, K., Seidel, T., and Gasteiger, J. (2007) Structure and reaction based evaluation of

synthetic accessibility. *J. Comput.-Aided Mol. Des.* **21**, 311–325.
156. Zaliani, A., Boda, K., Seidel, T., Herwig, A., Schwab, C. H., Gasteiger, J., Claussen, H., Lemmen, C., Degen, J., Paern, J., and Rarey, M. (2009) Second-generation de novo design: a view from a medicinal chemist perspective. *J. Comput.-Aided Mol. Des.* **23**, 593–602.
157. Patel, H., Bodkin, M. J., Chen, B., and Gillet, V. J. (2009) Knowledge-based approach to *de novo* design using reaction vectors. *J. Chem. Inf. Model.* **49**, 1163–1184.
158. Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534.
159. Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A., and Ringe, D. (1996) An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* **100**, 2605–2611.
160. Carr, R. A. E., Congreve, M., Murray, C. W., and Rees, D. C. (2005) Fragment-based lead discovery: leads by design. *Drug Discov. Today* **10**, 987–992.
161. Warr, W. (2009) Fragment-based drug discovery. *J. Comput.-Aided Mol. Des.* **23**, 453-458.
162. Joseph-McCarthy, D. (2009) Challenges of fragment screening. *J. Comput.-Aided Mol. Des.* **23**, 449–451.
163. Chen, I. J., and Hubbard, R. (2009) Lessons for fragment library design: analysis of output from multiple screening campaigns. *J. Comput.-Aided Mol. Des.* **23**, 603–620.
164. Hopkins, A. L., Groom, C. R., and Alex, A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **9**, 430–431.
165. Abad-Zapatero, C., and Metz James, T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* **10**, 464–469.
166. Leeson, P. D., and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **6**, 881–890.
167. Congreve, M., Chessari, G., Tisi, D., and Woodhead, A. J. (2008) Recent developments in fragment-based drug discovery. *J. Med. Chem.* **51**, 3661–3680.
168. Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003) A "rule of three" for fragment-based lead discovery? *Drug Discov. Today* **8**, 876–877.
169. Hajduk, P. J., and Greer, J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.* **6**, 211–219.
170. Erlanson, D. A., Wells, J. A., and Braisted, A. C. (2004) Tethering: fragment-based drug discovery. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 199–223, 194 plates.
171. Blomberg, N., Cosgrove, D., Kenny, P., and Kolmodin, K. (2009) Design of compound libraries for fragment screening. *J. Comput.-Aided Mol. Des.* **23**, 513–525.
172. de Kloe, G. E., Bailey, D., Leurs, R., and de Esch, I. J. P. (2009) Transforming fragments into candidates: small becomes big in medicinal chemistry. *Drug Discov. Today* **14**, 630–646.
173. Willett, P. (2009) Similarity methods in chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **43**, 3–71.
174. Eckert, H., and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **12**, 225–233.
175. Raymond, J. W., and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **16**, 521–533.
176. Ghose, A. K., Herbertz, T., Salvino, J. M., and Mallamo, J. P. (2006) Knowledge-based chemoinformatic approaches to drug discovery. *Drug Discov. Today* **11**, 1107–1114.
177. Accelrys. Pipeline Pilot http://accelrys.com/products/scitegic/ (accessed October 2, 2009).
178. Simulations Plus. ClassPharmer. http://www.simulations-plus.com/ (accessed October 6, 2009).
179. ChemAxon. Library MCS. http://www.chemaxon.com/shared/libMCS/ (accessed October 2, 2009).
180. Cramer, R. D. (2003) Topomer CoMFA: a design methodology for rapid lead optimization. *J. Med. Chem.* **46**, 374–388.
181. Jilek, R. J., and Cramer, R. D. (2004) Topomers: a validated protocol for their self-consistent generation. *J. Chem. Inf. Comput. Sci.* **44**, 1221–1227.
182. Cramer, R., and Wendt, B. (2007) Pushing the boundaries of 3D-QSAR. *J. Comput.-Aided Mol. Des.* **21**, 23–32.
183. Lill, M. A. (2007) Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* **12**, 1013–1017.
184. Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., and Duraiswami, C. (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **119**, 10509–10524.

185. Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen, X.-Q., Doweyko, A., and Li, Y. (2003) *In silico* ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **17**, 83–92.
186. Kubinyi, H. Why models fail. http://americanchemicalsociety.mediasite.com/acs/viewer/?peid=7a194d147-baa199-19-4b192d-a823-191fd117bf5301c (accessed September 22, 2009).
187. Tropsha, A., and Golbraikh, A. (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **13**, 3494–3504.
188. Hawkins, D. M. (2003) The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12.
189. Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **44**, 1912–1928.
190. Kubinyi, H., Hamprecht, F. A., and Mietzner, T. (1998) Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **41**, 2553–2564.
191. Kubinyi, H. (2006) Validation and predictivity of QSAR models, in *QSAR & Molecular Modelling in Rational Design of Bioactive Molecules, Proceedings of the 15th European Symposium on Structure-Activity Relationships and Molecular Modelling, Istanbul, Turkey 2004*, pp 30–33.
192. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., and Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **17**, 241–253.
193. Golbraikh, A., and Tropsha, A. (2002) Beware of q2! *J. Mol. Graphics Modell.* **20**, 269–276.
194. Doweyko, A. M. (2004) 3D-QSAR illusions. *J. Comput.-Aided Mol. Des.* **18**, 587–596.
195. Gramatica, P. (2007) Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **26**, 694–701.
196. Tropsha, A., Gramatica, P., and Gombar, V. K. (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Comb. Sci.* **22**, 69–77.
197. Jorgensen, W. L. (2006) QSAR/QSPR and proprietary data. *J. Chem. Inf. Model.* **46**, 937–937.
198. Editorial. (2006) QSAR/QSPR and proprietary data. *J. Med. Chem.* **49**, 3431–3431.
199. *ChemMedChem* notice to authors. http://www3.interscience.wiley.com/journal/110485305/home/110482452_notice.html?CRETRY=110485301&SRETRY=110485300 (accessed September 22, 2009).
200. Maggiora, G. M. (2006) On outliers and activity cliffs. Why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535–1535.
201. Peltason, L., and Bajorath, J. (2007) Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.* **14**, 489–497.
202. Peltason, L., and Bajorath, J. (2007) SAR Index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **50**, 5571–5578.
203. Guha, R., and Van Drie, J. H. (2008) Assessing how well a modeling protocol captures a structure-activity landscape. *J. Chem. Inf. Model.* **48**, 1716–1728.
204. Guha, R., and Van Drie, J. H. (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646–658.
205. Medina-Franco, J. L., Martinez-Mayorga, K., Bender, A., Marin, R. M., Giulianotti, M. A., Pinilla, C., and Houghten, R. A. (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **49**, 477–491.
206. Johnson, S. R. (2007) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **48**, 25–26.
207. Pavan, M., and Worth, A. P. Review of QSAR Models for Ready Biodegradation. http://ecb.jrc.ec.europa.eu/documents/QSAR/QSAR_Review_Biodegradation.pdf (accessed September 23, 2009).
208. Warr, W. A. Proceedings of New Horizons in Toxicity Prediction. Lhasa Limited symposium event in collaboration with the University of Cambridge, December 2008. http://www.qsarworld.com/files/Lhasa_Symposium_2008_Report.pdf (accessed September 23, 2009).
209. Huynh, L., Masereeuw, R., Friedberg, T., Ingelman-Sundberg, M., and Manivet, P. (2009) In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part I: beyond the reduction of animal model use. *Drug Discov. Today* **14**, 401–405.
210. Gramatica, P., Pilutti, P., and Papa, E. (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting

into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.* **44**, 1794–1802.
211. Hewitt, M., Cronin, M. T. D., Madden, J. C., Rowe, P. H., Johnson, C., Obi, A., and Enoch, S. J. (2007) Consensus QSAR models: do the benefits outweigh the complexity? *J. Chem. Inf. Model.* **47**, 1460–1468.
212. Matthews, E. J., Kruhlak, N. L., Benz, R. D., Contrera, J. F., Marchant, C. A., and Yang, C. (2008) Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows software to achieve high-performance, high-confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol. Mech. Methods* **18**, 189–206.
213. Abshear, T., Banik, G. M., D'Souza, M. L., Nedwed, K., and Peng, C. (2006) A model validation and consensus building environment. *SAR QSAR Environ. Res.* **17**, 311–321.
214. Boyer, S., Arnby, C. H., Carlsson, L., Smith, J., Stein, V., and Glen, R. C. (2007) Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inf. Model.* **47**, 583–590.
215. Terfloth, L., Bienfait, B., and Gasteiger, J. (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **47**, 1688–1701.
216. Molecular Networks. BioPath. http://www.mol-net.de/biopath/index.html (accessed September 25, 2009).
217. Schwaighofer, A., Schroeter, T., Mika, S., Hansen, K., ter Laak, A., Lienau, P., Reichel, A., Heinrich, N., and Müller, K.-R. (2008) A probabilistic approach to classifying metabolic stability. *J. Chem. Inf. Model.* **48**, 785–796.
218. Symyx Technologies. Metabolite. http://www.symyx.com/products/databases/bioactivity/metabolite/index.jsp (accessed October 2, 2009).
219. Ridder, L., and Wagener, M. (2008) SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* **3**, 821–832.
220. Button, W. G., Judson, P. N., Long, A., and Vessey, J. D. (2003) Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J. Chem. Inf. Comput. Sci.* **43**, 1371–1377.
221. Leach, A. R., Gillet, V. J., Lewis, R. A., and Taylor, R. (2010) Three-Dimensional Pharmacophore Methods in Drug Discovery. J. Med. Chem. **53**, 539–558
222. Wolber, G., Seidel, T., Bendix, F., and Langer, T. (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* **13**, 23–29.
223. Feng, J., Sanil, A., and Young, S. S. (2006) PharmID: pharmacophore identification using Gibbs sampling. *J. Chem. Inf. Model.* **46**, 1352–1359.
224. Richmond, N. J., Willett, P., and Clark, R. D. (2004) Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graphics Modell.* **23**, 199–209.
225. Richmond, N., Abrams, C., Wolohan, P., Abrahamian, E., Willett, P., and Clark, R. (2006) GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput.-Aided Mol. Des.* **20**, 567–587.
226. Jain, A. N. (2004) Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.* **47**, 947–961.
227. Cole, J. C., Gardiner, E. J., Gillet, V. J., and Taylor, R. (2009) Development of test systems for pharmacophore elucidation, in *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, March 22–26, 2009.*
228. Patel, Y., Gillet, V. J., Bravi, G., and Leach, A. R. (2002) A comparison of the pharmacophore identification programs: catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **16**, 653–681.
229. Clark, R. D. (2009) Prospective ligand- and target-based 3D QSAR: state of the art 2008. *Curr. Top. Med. Chem.* **9**, 791–810.
230. Deng, Z., Chuaqui, C., and Singh, J. (2003) Structural Interaction Fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding Interactions. *J. Med. Chem.* **47**, 337–344.
231. Muegge, I., and Oloff, S. (2006) Advances in virtual screening. *Drug Discov. Today: Technol.* **3**, 405–411.
232. Howe, T. J., Mahieu, G., Marichal, P., Tabruyn, T., and Vugts, P. (2006) Data reduction and representation in drug discovery. *Drug Discov. Today* **12**, 45–53.
233. Ivanenkov, Y. A., Savchuk, N. P., Ekins, S., and Balakin, K. V. (2009) Computational mapping tools for drug discovery. *Drug Discov. Today* **14**, 767–775.
234. Wagener, M., Sadowski, J., and Gasteiger, J. (1995) Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic AH receptor activity by neural networks. *J. Am. Chem. Soc.* **117**, 7769–7775.
235. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J., and Gasteiger, J. (1996) Locating biologically

active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **36**, 1205–1213.
236. Shneiderman, B. (1992) Tree visualization with tree-maps: 2-D space-filling approach. *ACM Trans. Graph.* **11**, 92–99.
237. Kibbey, C., and Calvet, A. (2005) Molecular property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.* **45**, 523–532.
238. Yamashita, F., Itoh, T., Hara, H., and Hashida, M. (2006) Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J. Chem. Inf. Model.* **46**, 1054–1059.
239. Lamping, J., Rao, R., and Pirolli, P. (1995) A focus + context technique based on hyperbolic geometry for visualizing large hierarchies, in *Proceedings of the SIGCHI conference on human factors in computing systems*, Denver, Colorado, United States, ACM Press/Addison-Wesley Publishing Co.
240. Agrafiotis, D. K., Bandyopadhyay, D., and Farnum, M. (2006) Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.* **47**, 69–75.
241. Agrafiotis, D. K., and Xu, H. (2002) A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. USA.* **99**, 15869–15872.
242. Agrafiotis, D. K., and Xu, H. (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **43**, 475–484.
243. Agrafiotis, D. K. (2003) Stochastic proximity embedding. *J. Comput. Chem.* **24**, 1215–1221.
244. Patel, A., Chin, D. N., Singh, J., and Denny, R. A. (2006) Methods for describing a group of chemical structures. WO 2006023574.
245. Medina-Franco, J. L., Petit, J., and Maggiora, G. M. (2006) Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem. Biol. Drug Des.* **67**, 395–408.
246. Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M. A., and Waldmann, H. (2006) The scaffold tree – visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **47**, 47–58.
247. Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893.
248. Lipkus, A. H., Yuan, Q., Lucas, K. A., Funk, S. A., Bartelt, W. F., Schenck, R. J., and Trippe, A. J. (2008) Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.* **73**, 4443–4451.
249. Schuffenhauer, A., Brown, N., Ertl, P., Jenkins, J. L., Selzer, P., and Hamon, J. (2007) Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.* **47**, 325–336.
250. Agrafiotis, D. K., Shemanarev, M., Connolly, P. J., Farnum, M., and Lobanov, V. S. (2007) SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.* **50**, 5926–5937.
251. Kolpak, J., Connolly, P. J., Lobanov, V. S., and Agrafiotis, D. K. (2009) Enhanced SAR Maps: Expanding the Data Rendering Capabilities of a Popular Medicinal Chemistry Tool. J. Chem. Inf. Model. **49**, 2221–2230.
252. Smellie, A. (2007) General purpose interactive physico-chemical property exploration. *J. Chem. Inf. Model.* **47**, 1182–1187.
253. Krallinger, M., Erhardt, R. A.-A., and Valencia, A. (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today* **10**, 439–445.
254. Ananiadou, S., Kell, D. B., and Tsujii, J.-I. (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.* **24**, 571–579.
255. Erhardt, R. A. A., Schneider, R., and Blaschke, C. (2006) Status of text mining techniques applied to biomedical text. *Drug Discov. Today* **11**, 315–325.
256. Banville, D. L. (2006) Mining chemical structural information from the drug literature. *Drug Discov. Today* **11**, 35–42.
257. Banville, D. L. (2009) Mining chemical and biological information from the drug literature. *Curr. Opin. Drug Discov. Dev.* **12**, 376–387.
258. Banville, D. L., (Ed.) (2009) *Chemical Information Mining: Facilitating Literature-Based Discovery*, CRC Press, Boca Raton, FL.
259. RSC Prospect. http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp (accessed October 2, 2009).
260. Batchelor, C. R., and Corbett, P. T. (2007) Semantic enrichment of journal articles using chemical named entity recognition, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* Prague, Czech Republic, Association for Computational Linguistics, http://www.aclweb.org/anthology/W/W07/W07-1008.pdf (accessed October 2, 2009)

261. Corbett, P., Batchelor, C., and Teufel, S. (2007) Annotation of chemical named entities, in *BioNLP 2007: Biological, translational, and clinical language processing*, Prague, Czech Republic, Association for Computational Linguistics, http://www.aclweb.org/anthology/W/W07/W07-1008.pdf (accessed October 2, 2009).
262. Kidd, R. Prospecting for chemistry in publishing. paper given at ICIC 2008, Nice France October 2008. http://www.infonortics.eu/chemical/ch08/slides/kidd.pdf (accessed October 2, 2008).
263. RSC Ontologies. http://www.rsc.org/ontologies/ (accessed October 2, 2009).
264. Valko, A. T., and Johnson, A. P. (2009) CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. *J. Chem. Inf. Model.* **49**, 780–787.
265. Eigner-Pitto, V., Eiblmaier, J., U. Frieske, U., Isenko, L., Kraut, H., Saller, H., and Loew, P. Mining for chemistry in text and images. A real world example revealing the challenge, scope, limitation and usability of the current technology. paper given at Fraunhofer-Symposium on Text Mining, Bonn, September 29–30, 2008. http://www.scai.fraunhofer.de/fileadmin/download/vortraege/tms_08/Valentina_Eigner_Pitto.pdf (accessed October 2, 2009).
266. Filippov, I. V., and Nicklaus, M. C. (2009) Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.* **49**, 740–743.
267. Park, J., Rosania, G., Shedden, K., Nguyen, M., Lyu, N., and Saitou, K. (2009) Automated extraction of chemical structure information from digital raster images. *Chem. Cent. J.* **3**, 4.
268. Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S. Y., Johnson, A. P., Major, S., Wade, R. A., and Ando, H. Y. (2009) Route Designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **49**, 593–602.
269. InfoChem. http://www.infochem.de (accessed October 2, 2009).
270. Matthews, E. J. Current Approaches for Toxicity Prediction, in *New Horizons in Toxicity Prediction. Lhasa Limited Symposium Event in Collaboration with the University of Cambridge.* http://www.qsarworld.com/lhasa_report1.php (accessed October 7, 2009).
271. Clark, R. D. (2009) At what point does docking morph into 3-D QSAR?, in *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, March 22–26, 2009.*
272. Scheiber, J., Chen, B., Milik, M., Sukuru, S. C. K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., Glick, M., Davies, J. W., and Jenkins, J. L. (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.* **49**, 308–317.
273. Schreiber, S. L., Kapoor, T. M., and Wess, G., (Eds.) (2007) *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, Wiley-VCH, Weinheim, Germany.

Chapter 2

Molecular Similarity Measures

Gerald M. Maggiora and Veerabahu Shanmugasundaram

Abstract

Molecular similarity is a pervasive concept in chemistry. It is essential to many aspects of chemical reasoning and analysis and is perhaps the fundamental assumption underlying medicinal chemistry. Dissimilarity, the complement of similarity, also plays a major role in a growing number of applications of molecular diversity in combinatorial chemistry, high-throughput screening, and related fields. How molecular information is represented, called the representation problem, is important to the type of molecular similarity analysis (MSA) that can be carried out in any given situation. In this work, four types of mathematical structure are used to represent molecular information: sets, graphs, vectors, and functions. Molecular similarity is a pairwise relationship that induces structure into sets of molecules, giving rise to the concept of chemical space. Although all three concepts – molecular similarity, molecular representation, and chemical space – are treated in this chapter, the emphasis is on molecular similarity measures. Similarity measures, also called similarity coefficients or indices, are functions that map pairs of compatible molecular representations that are of the same mathematical form into real numbers usually, but not always, lying on the unit interval. This chapter presents a somewhat pedagogical discussion of many types of molecular similarity measures, their strengths and limitations, and their relationship to one another. An expanded account of the material on chemical spaces presented in the first edition of this book is also provided. It includes a discussion of the topography of activity landscapes and the role that activity cliffs in these landscapes play in structure–activity studies.

Key words: Molecular similarity, Molecular similarity analyses, Dissimilarity, Activity landscapes

1. Introduction

Similarity is a fundamental concept that has been used since before the time of Aristotle. Even in the sciences, it has been used for more than two centuries [1]. Similarity is subjective and relies upon comparative judgments – there is no absolute standard of similarity, rather "like beauty, it is in the eye of the beholder." Because of this subjectivity, it is difficult to develop methods for unambiguously computing the similarities of large sets of

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_2,

© Springer Science+Business Media, LLC 2011

molecules [2]. Moreover, there is no absolute standard to compare to so that assessing the validity of any similarity-based method remains subjective; basically, one must rely upon the judgment of experienced scientists. Nevertheless, numerous approaches have been developed over the years to address this difficult but important problem [3–5].

The notion of similarity is fundamental to many aspects of chemical reasoning and analysis; indeed, it is perhaps the fundamental assumption underlying medicinal chemistry, and falls under the general rubric of *molecular similarity analysis* (MSA). Determining the similarity of one "molecular object" to another is basically an exercise in pattern matching – generally called the *matching problem*. The outcome of the exercise is a value, the *similarity measure* that characterizes the degree of matching, association, proximity, resemblance, alignment, or similarity of pairs of molecules as manifested by their "molecular patterns," which are made up of sets of features. The terminology "proximity" is sometimes used in a more general sense to refer to the similarity, dissimilarity, or distance between pairs of molecules. Similarity is generally considered to be a symmetric property, that is "A" is as similar to "B" as "B" is to "A," and most studies are based upon this property. Tversky [6], however, has argued persuasively that certain similarity comparisons are inherently asymmetric. Although his work was directed towards psychology it nonetheless has applicability in studies of molecular similarity. An example will be presented that illustrates the nature of asymmetric similarity and how it can be used to augment the usefulness of the usual symmetric version of similarity (*Cf.* the discussion of Chen and Brown [7]). Willett, Barnard, and Downs [8] presented a comprehensive overview of many of the similarity measures in use today. Their review included a table that summarized the form of the various measures with respect to the type of representation used and should be consulted for further details. Bender and Glen [9] have provided a more recent review of molecular similarity.

Choosing an appropriate feature set and an associated mathematical structure (e.g. set, vector, function, or graph) for handling them is called the *representation problem* and underlies all aspects of MSA. Because similarity is subjective, choosing a feature set depends upon the background of the scientist doing the choosing and to some extent on the problem being addressed. For example, a synthetic organic chemist may focus on the nature of a molecular scaffold and its substituent groups while a physical chemist may be more interested in 3-D shape and electrostatic properties.

Closely allied with the notion of molecular similarity is that of a *chemical space*. Chemical spaces provide a means for conceptualizing and visualizing the molecular similarities of large sets of molecules. A chemical space consists of a set of molecules and a set of associated relations (e.g. similarities, dissimilarities, distances,

etc.) among the molecules that give the space a "structure" [10]. In most chemical spaces, which are coordinate-based, molecules are generally depicted as points. This, however, need not always be the case – sometimes only similarities or "distances" among molecules in the population are known. Nevertheless, this type of pairwise information can be used to construct appropriate coordinate systems using methods such as multi-dimensional scaling (MDS) [11], principal-component analysis (PCA) [12], or nonlinear mapping (NLM) [13] that optimally preserve the information. Coordinate-based chemical spaces can also be partitioned into cells and are usually referred to as cell-based chemical spaces [14]. Each particular type of representation of chemical space has its strengths and weaknesses so that it may be necessary to use multiple types of representations to satisfactorily treat specific problems.

Identifying the appropriate molecular features is crucial in MSA, since the number of potential features is quite large and many contain redundant information. Typical types of molecular features include molecular size, shape, charge distribution, conformation states, and conformational flexibility. In general, only those features deemed relevant or necessary to the matching task at hand are considered. Features are mimicked by any number of descriptors that, ideally, capture the essential characteristics of the features. For example, numerous descriptors of molecular shape exist such as the Jurs shape indices [15] or the Sterimol parameters [16] as well as descriptors of charge distributions such as the venerable Mulliken population analysis [17] or charged partial surface areas, which conveniently incorporate both charge and shape information [18] and descriptors of conformational flexibility such as the Kier molecular flexibility index Φ [19]. Sometimes the term "feature" is used interchangeably with "descriptor." As is seen in the above discussion, features are more general than descriptors, but this distinction is generally not strictly adhered to in most research papers including this one. Other chapters in this work should be consulted for detailed discussion of the many types and flavors of descriptors in use in cheminformatics and chemometrics today.

Similarity measures for assessing the degree of matching between two molecules given a particular representation constitute the main subject matter of this chapter. These measures are functions that map pairs of *compatible* molecular representations (i.e. representations of the same mathematical form) into real numbers usually, but not always, lying on the unit interval. Set, graph, vector, and function-based representations employ a variety of distance and "overlap" measures. Graph-based representations use chemical distance or related graph metrics [20, 21], although numerous graph invariants have been employed as descriptors in vector-based representations [22–24]. All of the

similarity and related measures have at least some idiosyncratic behavior, which can give rise to misleading assessments of similarity or dissimilarity [2]. Similarity measures are sometimes referred to as similarity coefficients or similarity indices and these terminologies will be used somewhat interchangeably in this work.

From the above discussion it is clear that similarity measures provide assessments that are inherently subjective in nature. Thus, the inconsistencies of various measures are not entirely surprising and sometimes can be quite daunting. An interesting approach was developed by Willett's group using a technique called "data fusion" [25]. They showed that values obtained from multiple similarity methods combined using data fusion led to an improvement over similarity-based compound searching using a single similarity method. Subsequent work by Willett's group extended the methodology [26, 27] and provided a detailed account of its conceptual framework [28]. Alternatively, less sophisticated, approaches such as taking the mean of multiple similarity values can also be used.

A brief introduction to the types of molecular representations typically encountered in MSA is presented at the beginning of Subsection 2 followed in Subsection 2.1 by a discussion of similarity measures based upon chemical-graph representations. While graph-based representations are the most familiar to chemists, their use has been somewhat limited in similarity studies due to the difficulty of evaluating the appropriate similarity measures. This section is followed by a discussion of similarity measures based upon finite vector representations, the most ubiquitous types of representations. In these cases, the vector components can be of four types

$$\begin{aligned} &\mathbb{B} && \text{Boolean Variables } \{0,1\} \\ &\mathbb{K} && \text{Categorical Variables } \{\text{finite, ordered set}\} \\ &\mathbb{N}_o && \text{Non - Negative Integer Variables } \{0,1,2,3,\ldots\} \\ &\mathbb{R} && \text{Real Variables } \{\text{uncountably infinite set}\} \end{aligned} \tag{1}$$

the first of which called "binary vectors," "bit vectors," or "molecular fingerprints" is by far the most prevalent in applications and is discussed in detail in Subsection 2.2.1. Although the terminology "vector" is used, these objects mathematically are classical sets. Thus, the associated similarity measures are set-based rather than vector-based measures. In addition to the more traditional symmetric similarity measures, a discussion of *asymmetric* similarity measures associated with binary vectors is presented in Subsection 2.2.2.

Vectors whose components are based upon categorical or integer variables are described in Subsection 2.2.3. As was the case for binary vectors, these vectors are also classical sets, and as was the case in the previous subsection, the associated similarity

measures are set-based rather than vector-based. Here, it will also be seen that the form of the set measures are, in some cases, modified from those associated with traditional classical sets.

Subsection 2.3 describes the last class of finite feature vectors, namely those with continuous-valued components, where the components (i.e. features) are usually obtained from computed or experimentally measured properties. An often-overlooked aspect of continuous feature vectors is the inherent non-orthogonality of the basis of the "feature space." The consequences of this are discussed in Subsection 2.3.2. Similarity measures derived from continuous vectors are generally related to Euclidean distances or to cosine or correlation coefficients, all of which are vector-based measures, and are discussed in Subsection 2.3.3.

Essentially, none of the previously discussed approaches deals with the three-dimensionality of molecules. This is dealt with in Subsection 2.4, which describes the application of field-based functions to 3-D molecular similarity. The fields referred to here are related to the steric, electrostatic, and lipophilic properties of molecules and are represented by functions (i.e. "infinite-dimensional vectors"), which are usually taken to be linear combinations of atomic-centered Gaussians. Similarity measures totally analogous to those defined for finite-dimensional, continuous-valued feature vectors (*see* Subsection 2.3.3) also apply here and are treated in Subsection 2.4.2. An often unappreciated issue in 3-D molecular similarity studies is that of *consistent multi-molecule 3-D alignments*, which are discussed in Subsection 2.4.3. Consider the alignment of molecules A and B and that of molecules B and C. Superimposing the aligned pairs using molecule B as a reference induces an alignment of molecules A and C. Now align molecules A and C independently. A consistent multi-molecule alignment is one in which both the induced and independent alignments are essentially the same. As was discussed by Mestres et al. [29], this approach is helpful in identifying "experimentally correct" alignments for a set of reverse transcriptase inhibitors even though the proper conformer of one of the molecules was not in its computed lowest-energy conformation. The role of conformational flexibility in 3-D MSA is discussed in general terms in Subsection 2.4.4. Two general approaches to this problem are described here. One involves the identification of a set of conformational prototypes and the other involves the simultaneous maximization of the similarity measure and minimization of the conformational energy of the molecules being aligned. The former approach is more computationally demanding because it involves $M \times N$ pairwise comparisons, where M and N are the respective numbers of prototype conformations for each pair of molecules. Given that multiple conformations may be important in MSA, how does one determine a similarity value that accounts for multiple conformational states? The discussion in Subsection 2.4.5 suggests an

approach that employs a weighting function based on Boltzmann-like probabilities for each of the conformational states.

Subsection 2.5 provides a brief discussion of molecular dissimilarity, a subject of importance when considering a variety of topics from selecting diverse subsets from a compound collection to the design of diverse combinatorial compound libraries.

The emerging role of *chemical space* in cheminformatics is treated in Subsection 3. It includes a discussion of the dimension of chemical spaces in Subsection 3.1 and a description in Subsection 3.2 of the methods for constructing coordinate-based and coordinate-free chemical spaces, how they can be transformed into one another, and how the usually high-dimension of typical chemical spaces can be reduced in order to facilitate visualization and analysis. The closely related subject of *activity cliffs* and the topography of activity landscapes are discussed in Subsection 3.3. How the information contained in activity landscapes, which are inherently of high-dimension, can be portrayed in lower dimensions is discussed. Emphasis is placed on the use of structure–activity similarity (SAS) maps, although several other recent approaches to this problem are described. An information-theoretic analysis of SAS maps is also presented. Subsection 3 ends with a somewhat detailed description of a general similarity-based approach for representing chemical spaces (*see* Subsection 3.4). The method explicitly accounts for the inherent non-orthogonality of vector representations of chemical space. Unlike some of the vector-like methods described earlier, this method employs "molecular vectors" that actually live in a linear vector space.

The present work is not intended as a comprehensive review of the similarity literature. Rather, it is intended to provide an integrated and somewhat pedagogical discussion of many of the simple, complex, and confounding issues confronting scientists using the concept of molecular similarity in their work.

2. Molecular Representations and Their Similarity Measures

How the structural information in molecules is represented is crucial to the types of "chemical questions" that can be asked and answered. This is certainly true in MSA where different representations and their corresponding similarity measures can lead to dramatically different results [2]. Four types of mathematical objects are typically used to represent molecules – sets, graphs, vectors, and functions. Sets are the most general objects and basically underlie the other three and are useful in their own right as will be seen below. Because of their importance a brief introduction to sets, employing a more powerful but less familiar

notation than that typically used, is provided in the Appendix (*see* Subsection 5).

Typically, chemists represent molecules as "chemical graphs" [30], which are closely related to the types of graphs dealt with by mathematicians in the field of graph theory [31]. Most chemical graphs describe the nature of the atoms and how they are bonded. Thus, chemical graphs are sometimes said to provide a 2-D representation of molecules. They do not typically contain information on the essential 3-D features of molecules, although chemical graphs have been defined that do capture some of this information [32]. Three-dimensional structures are also used extensively, especially now that numerous computer programs have been developed for their computation and display.

While chemical graphs provide a powerful and intuitive metaphor for understanding many aspects of chemistry, they nevertheless have their limitations especially when dealing with questions of interest in chemometrics and cheminformatics. In these fields, molecular information is typically represented by *feature vectors*, where each component corresponds to a "local" or "global" feature or property of a molecule usually represented by one of a number of possible descriptors associated with the chosen feature. Local features include molecular fragments ("substructures"), potential pharmacophores [33], various topological indices [34], and partial atomic charges, to name a few. Global features include properties such as molecular weight, logP, polar surface area, various BCUTs [35], and volume. It is well to point out here that use of the term "vector" is not strictly correct. For example, in Subsection 2.2 on Discrete-Valued Feature Vectors the "bit vectors" used to depict molecular fingerprints are actually classical sets or multisets that do not strictly behave according to the rules of linear vector spaces [36]. For example, bit vectors $\mathbf{v}_A$ and $\mathbf{v}_B$ do not satisfy the additive ("+") and scalar multiplicative (" $\cdot$ ") properties associated with vectors residing in linear vector spaces, i.e. $\mathbf{v}_C \neq a \cdot \mathbf{v}_A + b \cdot \mathbf{v}_B$, where a and b are any real scalars. As discussed in the sequel, some classes of continuous-valued vectors (*see* Subsection 2.3) such as BCUTS are also not vectors in the strict mathematical sense since they do not strictly obey the additive property of vectors (e.g. the sum of two BCUTS may lie outside of the BCUT chemical space).

More recently, with the significant increases in computer power even on desktop PCs, methods for *directly matching* 3-D features of molecules have become more prevalent. Features here generally refer to various types of molecular fields, some such as electron density ("steric") and electrostatic-potential fields are derived from fundamental physics [37, 38] while others such as lipophilic potential fields [39] are constructed in an ad hoc manner. Molecular fields are typically represented as continuous functions. As is the case for discrete and continuous vectors noted in

the previous paragraph, such functions also may not strictly satisfy the axioms of linear function spaces [40]. However, this does not preclude their usefulness in determining 3-D molecular similarities. Discrete fields have also been used [41], albeit somewhat less frequently except in the case of the many CoMFA-based studies [42].

In cheminformatics, similarities typically are taken to be real numbers that lie on the unit interval [0,1]. However, since similarity is an inherently "fuzzy" concept, it may be appropriate to take a fuzzier view. Bandemere and Näther [43] have written extensively on an approach that treats similarity as a fuzzy number [44]. Fuzzy numbers can be conceptualized as Gaussian functions or, more commonly as "teepee-like" functions, with maximum value unity, the "width" of the function being associated with the "degree of fuzziness" of the numbers. While this is a conceptually powerful approach, it suffers many of the problems associated with fuzzy numbers. For example, it two fuzzy numbers overlap significantly determining how much larger one fuzzy number is to the other can become difficult [44]. Thus, for example, comparing how similar two molecules are to a given molecule can become a significant problem. Nevertheless, investigating the realm of fuzzy similarities may prove to be a suitable approach to similarity, but further work needs to be done before any reasonable conclusion can be obtained as to its usefulness in cheminformatics.

2.1. Chemical Graphs

Chemical graphs are ubiquitous in chemistry. A chemical graph, G_k, can be defined as an ordered triple of sets

$$G_k = (V_k,\ E_k,\ L_k), \tag{2}$$

where V_k is a set of n vertices ("atoms") and $V_k(x_i)$ is an *indicator* or *characteristic function* with values $V_k(x_i) = \{0, 1\}$ that designates the respective absence or presence of a given vertex

$$V_k = \{V_k(x_1), V_k(x_2), ..., V_k(x_n)\}. \tag{3}$$

Alternatively, V_k can be given, in more familiar notation, by

$$V_k = \{v_{k,k_1}, v_{k,k_2}, ..., v_{k,k_\ell}\} \tag{4}$$

where only the ℓ vertices for which $V(x_i) = 1$ are explicitly designated (*See* Appendix Subsection 5 for notational details). The edge set, E_k, can be written in analogous fashion,

$$E_k = \{e_{k,k_1}, e_{k,k_2}, ..., e_{k,k_m}\}, \tag{5}$$

where the m elements of the set are the edges ("bonds") between vertices ("atoms"), and each edge is associated with an unordered pair of vertices, $e_{k,k_i} = \{v_{k,k_p}, v_{k,k_q}\}$. The label set, L_k, is a set of r symbols

$$L_k = \{\ell_{k,k_1}, \ell_{k,k_2}, ..., \ell_{k,k_r}\} \tag{6}$$

that label each vertex ("atom") and/or edge ("bond"). Typical atom labels include hydrogen ("H"), carbon ("C"), nitrogen ("N"), and oxygen ("O"); typical bond labels include single ("s"), double ("d"), triple ("t"), and aromatic ("ar"), but other possibilities exist. Whatever symbol set is chosen will depend to some degree on the nature of the problem being addressed. In most cheminformatics applications, *hydrogen suppressed* chemical graphs are used, which are obtained by deleting all of the hydrogen atoms. Fig. 1 depicts an example of two hydrogen-suppressed chemical graphs, G_1 and G_2, which are clearly related to a chemist's 2-D representation of a molecule. Chemical graphs of 3-D molecular structures are described by Raymond and Willett [32], but their use has been much more limited.

The notion of a subgraph is also important. If G'_k is a subgraph of G_k, written $G'_k \subseteq G_k$, then

$$G'_k \subseteq G_k \Rightarrow V'_k \subseteq V_k \text{ and } E'_k \subseteq E_k, \tag{7}$$

that is the vertex and edge sets V'_k and E'_k associated with the subgraph, G'_k, are subsets of the corresponding vertex and edge sets V_k and E_k of the graph, G_k. Many operations defined on sets

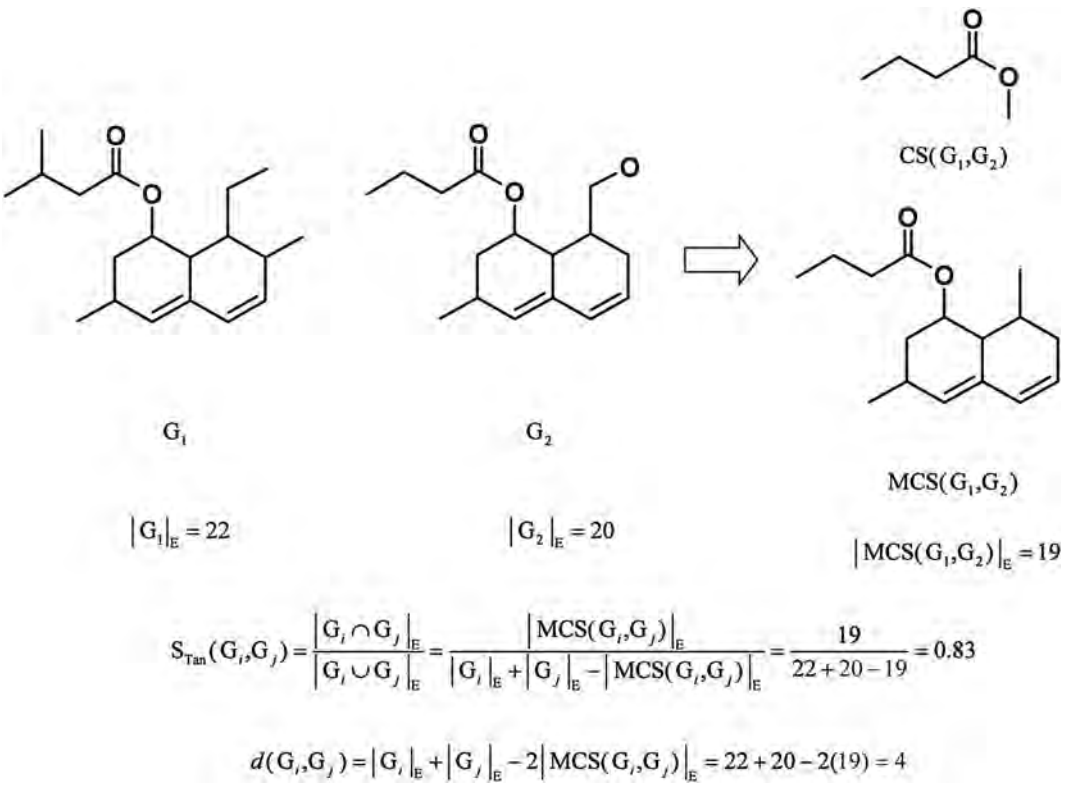


Fig. 1. An example of two hydrogen-suppressed graphs G1, G2 and a common substructure CS(G1,G2) and the maximum common substructure MCS(G1,G2) are shown above. The Tanimoto similarity index and the distance between the two chemical graphs are computed below.

can also be defined on graphs. One such operation is the norm or cardinality of a graph,

$$|G_k| = |V_k| + |E_k| \tag{8}$$

which is a measure of the "size" of the graph. Another measure is the *edge norm* that is given by

$$|G_k|_{\mathrm{E}} = |E_k|, \tag{9}$$

where the subscript E explicitly denotes that the cardinality refers only to the edges ("bonds") of the graph. For the two chemical graphs depicted in Fig. 1, $|G_1|_{\mathrm{E}} = 22$ and $|G_2|_{\mathrm{E}} = 20$. Note that only the number of bonds and not their multiplicities (e.g. single, double, etc.) are considered here. However, many other possibilities exist, and their use will depend on the problem being addressed [20].

A key concept in the assessment of molecular similarity based upon chemical graphs is that of a *maximum common substructure*, $\mathrm{MCS}(G_i, G_j)$, of two chemical graphs, which derives from the concept of maximum common subgraph employed in mathematical graph theory. There are several possible forms of MCS [21, 32]. Here, we will focus on what is usually called the maximum common edge substructure, which is closest to what chemists perceive as "chemically meaningful" substructures [45], but we will retain the simpler and more common nomenclature MCS. A common (edge) substructure (CS) of two chemical graphs is given by

$$\mathrm{CS}(G_i, G_j)_{k,\ell} = E_i^k \cap E_j^\ell = E_i^k = E_j^\ell, \tag{10}$$

where E_i^k and E_j^ℓ are subsets of their respective edge sets, $E_i^k \subseteq E_i$ and $E_j^\ell \subseteq E_j$, and are equivalent. Thus, the intersection (or union) of these two equivalent subsets is equal to the sets themselves. As there are numerous such common substructures, $\mathrm{CS}(G_i, G_j)_{k,\ell}$, $k, \ell = 1, 2, 3, \ldots\ldots$, determining the MCS between two chemical graphs is equivalent to determining the edge intersection-set of maximum cardinality, that is

$$\begin{aligned} \mathrm{MCS}(G_i, G_j) &= \mathrm{CS}(G_i, G_j)_{p,q} \text{ such that } \left|\mathrm{CS}(G_i, G_j)_{p,q}\right|_{\mathrm{E}} \\ &= \max_{k,\ell} \left|\mathrm{CS}(G_i, G_j)_{k,\ell}\right|_{\mathrm{E}} \end{aligned} \tag{11}$$

Thus,

$$G_i \cap G_j \equiv \mathrm{MCS}(G_i, G_j), \tag{12}$$

that is the MCS is equivalent to "graph intersection," which is equivalent to the maximum number of edges in common between the two molecules. Note that multiple solutions may exist and that some of the solutions could involve disconnected graphs.

However, to obtain "chemically meaningful" results only *connected* MCS's are usually considered.

The edge cardinality of the intersection and union of two chemical graphs is given, respectively, by

$$\left|G_i \cap G_j\right|_{\mathrm{E}} = \left|\mathrm{MCS}(G_i, G_j)\right| \tag{13}$$

and

$$\left|G_i \cup G_j\right|_{\mathrm{E}} = \left|G_i\right|_{\mathrm{E}} + \left|G_j\right|_{\mathrm{E}} - \left|\mathrm{MCS}(G_i, G_j)\right|. \tag{14}$$

These two expressions form the basis for several measures such as Tanimoto similarity (*see* Subsection 2.2 for an extensive discussion)

$$S_{\mathrm{Tan}}(G_i,G_j) = \frac{\left|G_i \cap G_j\right|_{\mathrm{E}}}{\left|G_i \cup G_j\right|_{\mathrm{E}}} = \frac{\left|\mathrm{MCS}(G_i,G_j)\right|_{\mathrm{E}}}{\left|G_i\right|_{\mathrm{E}} + \left|G_j\right|_{\mathrm{E}} - \left|\mathrm{MCS}(G_i,G_j)\right|_{\mathrm{E}}} \tag{15}$$

and the distance between two chemical graphs

$$d(G_i,G_j) = \left|G_i\right|_{\mathrm{E}} + \left|G_j\right|_{\mathrm{E}} - 2\left|\mathrm{MCS}(G_i,G_j)\right|_{\mathrm{E}}. \tag{16}$$

The edge cardinality is explicitly designated in Eqs. (11) and (13)–(16) in order to emphasize that a particular norm has been chosen. Equation (15) is the graph-theoretical analog of the well-known Tanimoto similarity index (*see* Eq. (21)), which is symmetric and bounded by zero and unity. Equation (16) corresponds to the distance between two graphs [46], which is the number of bonds that are not in common in the two molecules depicted by G_i and G_j. Another distance measure called "chemical distance" is similar to that given in Eq. (16) except that lone-pair electrons are explicitly accounted for [21]. The Tanimoto similarity index of the two chemical graphs in Fig. 1 and the distance between them are given by $S_{\mathrm{Tan}}(G_i,G_j) = 0.83$ and $d(G_i,G_j) = 4$, respectively.

A similarity index called "subsimilarity," which is short for substructure similarity, has been developed by developed by Hagadone [47]. In form it is identical to one of the family of asymmetric similarity indices developed by Tversky [6] that is discussed in Subsection 2.2.2,

$$S_{\mathrm{Tev}}(G_{\mathrm{Q}},G_{\mathrm{T}}) = \frac{\left|G_{\mathrm{Q}} \cap G_{\mathrm{T}}\right|_{\mathrm{E}}}{\left|G_{\mathrm{Q}}\right|_{\mathrm{E}}} = \frac{\left|\mathrm{MCS}(G_{\mathrm{Q}},G_{\mathrm{T}})\right|_{\mathrm{E}}}{\left|G_{\mathrm{Q}}\right|_{\mathrm{E}}}, \tag{17}$$

where $G_{\mathrm{Q}}$is the substructure query and G_{T} is a target molecule. In contrast to $S_{\mathrm{Tan}}(G_i,G_j)$, $S_{\mathrm{Tev}}(G_i,G_j)$ is not symmetric, although zero and unity also bound it.

While chemical graphs are intuitive to those trained in the chemical sciences, they have not been widely used in MSA primarily because of the computational demands brought on by the need to compute $\mathrm{MCS}(G_i,G_j)$, which for large complex systems can be quite daunting. Approximate algorithms do exist, however

[32, 47] and with the ever-increasing power of computers the use of graph-based similarity may become more prevalent in the future. Interestingly, there is a close analogy between determination of the MCS and alignment of the 3-D molecular fields of molecules (*see* Subsection 2.4) except that in the former the optimization is discrete while in the latter it is continuous.

2.2. Discrete-Valued Feature Vectors

The components of discrete feature vectors may indicate the presence or absence of a feature, the number of occurrences of a feature, or a finite set of binned values such as would be found in an ordered, categorical variable.

2.2.1. Binary-Valued Feature Vectors

Each component of an n-component binary feature vector, also called *bit vectors* or *molecular fingerprints*,

$$\mathbf{v}_A = (v_A(x_1), v_A(x_2), ..., v_A(x_k), ..., v_A(x_n)) \tag{18}$$

indicates the presence or absence of a given feature, x_k, that is

$$v_A(x_k) = \begin{cases} 1 & \text{Feature present} \\ 0 & \text{Feature absent} \end{cases}. \tag{19}$$

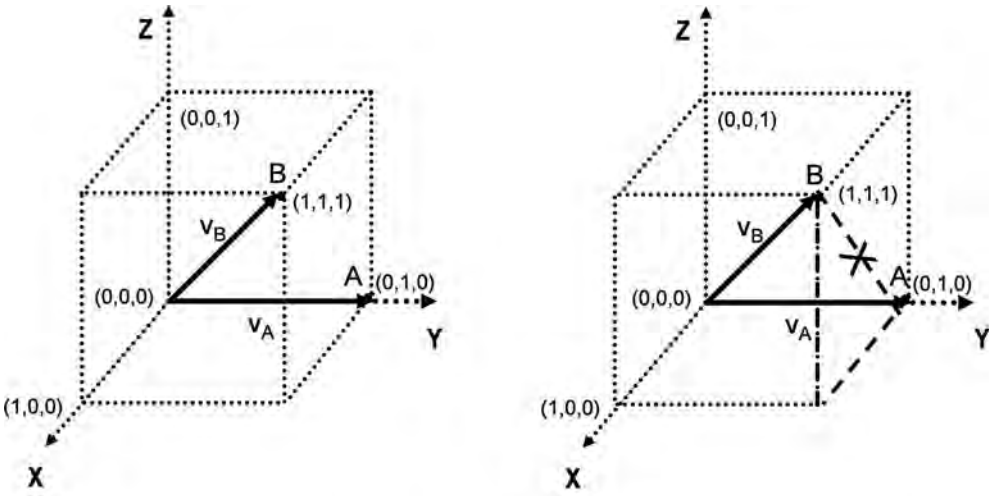
A wide variety of features have been used in bit vectors. These include molecular fragments, 3-D "potential pharmacophores," atom pairs, 2-D pharmacophores, topological torsions, and variety of topological indices to name a few.

Binary feature vectors are completely equivalent to sets (*see* the Appendix in Subsection 5 for further discussion). Care must be exercised when using them to ensure that appropriate mathematical operations are carried out. The number of components in a bit vector is usually quite large, normally $n >> 100$. In some cases n can be orders of magnitude larger, sometimes exceeding a million components [33, 48]. Bit vectors of this size are not handled directly since many of the components are zero, and methods such as hashing [49] are used to reduce the size of the stored information.

Bit vectors live in an n-dimensional, discrete hypercubic space, where each vertex of the hypercube corresponds to a set. Figure 2 provides an example of sets with three elements. Distances between two bit vectors, $\mathbf{v}_A$ and $\mathbf{v}_B$, measured in this space correspond to Hamming distances, which are based upon the city-block ℓ_1 metric

$$d_{Ham}(\mathbf{v}_A, \mathbf{v}_B) = |\mathbf{v}_A - \mathbf{v}_B| = \sum_{k=1}^{n} |v_A(x_k) - v_B(x_k)|. \tag{20}$$

Since these vectors live in an n-dimensional hypercubic space, the use of non-integer distance measures is inappropriate, although in this special case the square of the Euclidean distance is equal to the Hamming distance.



$$d_{\text{Ham}}(\mathbf{v}_A, \mathbf{v}_B) = |\mathbf{v}_A - \mathbf{v}_B| = \sum_{k=1}^{n} |v_A(x_k) - v_B(x_k)| = [1-0] + [1-1] + [1-0] = 2$$

Fig. 2. Distance between two binary-valued feature vectors $\mathbf{v}_A$ and $\mathbf{v}_B$ is not given by the Euclidean distance but the Hamming distance between the two.

The most widely used similarity measure by far is the Tanimoto similarity coefficient S_{Tan}, which is given in set-theoretic language as (*Cf.* Eq. (15) for the graph-theoretical case)

$$S_{\text{Tan}}(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{21}$$

Using the explicit expressions for set cardinality, intersection and union given by Eqs. (116, 111 and 112), respectively, in the Appendix, Eq. (21) becomes

$$S_{\text{Tan}}(A, B) = \frac{\sum_k \min[A(x_k), B(x_k)]}{\sum_k \max[A(x_k), B(x_k)]}. \tag{22}$$

By changing the form of the denominator (*see* Eqs. (120) and (121)), S_{Tan} is also given by

$$\begin{aligned} S_{\text{Tan}}(A, B) &= \frac{|A \cap B|}{|A - B| + |B - A| + |A \cap B|}, \\ &= \frac{a}{a + b + c} \end{aligned} \tag{23}$$

where

$$\begin{aligned} a &= |A \cap B| \quad \text{Number of features common to } A \text{ and } B \\ b &= |A - B| \quad \text{Number of features common to } A \text{ but not to } B. \\ c &= |B - A| \quad \text{Number of features common to } B \text{ but not to } A \end{aligned} \tag{24}$$

The Tanimoto similarity coefficient is symmetric,

$$S_{\mathrm{Tan}}(A,\ B) = S_{\mathrm{Tan}}(B,\ A), \tag{25}$$

as are most of the similarity coefficients in use today, and is bounded by zero and unity,

$$0 \leqslant S_{\mathrm{Tan}}(A,\ B) \leqslant 1. \tag{26}$$

From the form of these equations it can be seen that the method is biased when there is a great disparity in the size of the two molecules being compared. Consider, for example, the case when $|Q| << |T|$, where Q is a query molecule and T is a target molecule that could be obtained in a similarity search. If Q is much smaller than T, $|Q \cup T| \approx |T|$, and since $|Q| \leqslant |Q \cap T|$, it follows that $S_{\mathrm{Tan}}(Q, T) \approx |Q|/|T|$. A consequence of this relationship is that in similarity-based searching Q will tend to recover other small molecules, T, since as T gets larger S_{Tan} becomes smaller in value, which works against the selection of larger molecules in the search. This is not generally a problem except in cases where a substructure of a large target molecule is quite similar to the smaller query molecule. If the query were biologically active, the larger target molecule containing a similar substructure to the query, which is bioactive, would be missed. The same holds true for a large molecule query that is it will tend to recover larger molecules. Thus, molecules with a strong *substructural relationship* to the query molecule will likely be missed, but this could be important in drug design as the substructure may contain the key atoms of the pharmacophore. As will be seen in the next section, the use of an asymmetric similarity measure can compensate for this to some degree. The above argument carries through completely to the case of chemical-graph-based similarity indices (*see* Subsection 2.1).

A number of other similarity indices are in use today. The recent work by Willett, Barnard, and Downs [8] should be consulted for examples of many of them including a comprehensive discussion of their properties.

2.2.2. Asymmetric Similarity Indices

Most similarity measures for binary-valued feature vectors in use today are symmetric, Tversky [6], however, has defined an infinite family of *asymmetric* measures

$$S_{\mathrm{Tve}}(A, B) = \frac{|A \cap B|}{\alpha|A - B| + \beta|B - A| + |A \cap B|}, \tag{27}$$

where $\alpha, \beta \geqslant 0$. This generalizes the typical symmetric Tanimoto similarity measure given in Eq. (23), which obtains when $\alpha = \beta = 1$. For all other values of α and β, $S_{\mathrm{Tve}}(A, B)$is asymmetric, that is $S_{\mathrm{Tve}}(A, B) \neq S_{\mathrm{Tve}}(B, A)$. Only the two extreme forms will,

however, be considered here, namely those when $\alpha = 1$ and $\beta = 0$ and $\alpha = 0$ and $\beta = 1$. Their set-theoretic forms are given by

$$S^*_{\text{Tve}}(A, B) = \frac{|A \cap B|}{|A - B| + |A \cap B|} = \frac{|A \cap B|}{|A|} \quad \text{Fraction of } A \text{ similar to } B \tag{28}$$

$$S^*_{\text{Tve}}(B, A) = \frac{|A \cap B|}{|B - A| + |A \cap B|} = \frac{|A \cap B|}{|B|} \quad \text{Fraction of } B \text{ similar } A \tag{29}$$

Using Eqs. (111) and (116) both of the above equations can be written in a form similar to that for S_{Tan} given in Eq. (22). For example, Eq. (28) becomes

$$S^*_{\text{Tve}}(A, B) = \frac{\sum_k \min[A(x_k), B(x_k)]}{\sum_k A(x_k)}. \tag{30}$$

In analogy to Eq. (23), the asymmetric similarity indices are given, respectively, by

$$S^*_{\text{Tve}}(A, B) = \frac{a}{a + b} \text{ and } S^*_{\text{Tve}}(B, A) = \frac{a}{a + c}. \tag{31}$$

As was the case for the symmetric similarity coefficient

$$0 \leqslant S^*_{\text{Tve}}(A, B), S^*_{\text{Tve}}(B, A) \leqslant 1, \tag{32}$$

although $S^*_{\text{Tev}}(A, B) \neq S^*_{\text{Tev}}(B, A)$, in general. If $|A| < |B| \Rightarrow S^*_{\text{Tev}}(A, B) > S^*_{\text{Tev}}(B,A)$.

Asymmetric similarity can provide some benefits in similarity searches not afforded by its symmetric competitors. For example, consider as in Subsection 2.2.1, the query and target molecules, Q and T, respectively, and the asymmetric similarity coefficients given in Eqs. (28) and (29). If Q is relatively "small," (*N.B.* "small" and "large" are used here refer to the size of the set and not to the size of the corresponding molecule) that is if $|Q| << |T|$, then target molecules for which Q is an approximate subset will be selected using Eq. (28), that is

$$S^*_{\text{Tve}}(Q, T) = \frac{|Q \cap T|}{|Q|} \Rightarrow 1 \text{ as } Q \cap T \Rightarrow Q. \tag{33}$$

This result is approximately independent of the size of T given that Q is an approximate subset of T. A comparable selection of molecules would not be obtained using the symmetric similarity coefficient in Eq. (21) or the asymmetric similarity coefficient given by Eq. (29) since as the target molecule increased in size the denominator would reduce the overall similarity values

making selection less likely. If, on the other hand, Q is a relatively "large," that is if $|Q|>>|T|$, then using the lower expression for asymmetric similarity in Eq. (29) will produce similar results

$$S^{*}_{\mathrm{Tve}}(T, Q) = \frac{|Q \cap T|}{|T|} \Rightarrow 1 \text{ as } Q \cap T \Rightarrow T \tag{34}$$

except that the target molecules retrieved will be smaller than Q and will also be approximate subsets of Q. An example of this is shown in Figs. 3 and 4.

The "extreme" forms, but not the intermediate forms, of asymmetric similarity defined by Tversky [6] given in Eqs. (28) and (29) can be transformed into two symmetric measures by taking the maximum and minimum of the set cardinalities in the denominators of the two equations. The forms of these equations are obtained in analogy to those developed by Petke [41] for vectors and field-based functions (*see* Subsections 2.3 and 2.4 for further details):

$$S_{\mathrm{Pet}_{\max}}(A,\ B) = \frac{|A \cap B|}{\max(|A|, |B|)} \tag{35}$$

and

$$S_{\mathrm{Pet}_{\min}}(A,\ B) = \frac{|A \cap B|}{\min(|A|, |B|)}. \tag{36}$$

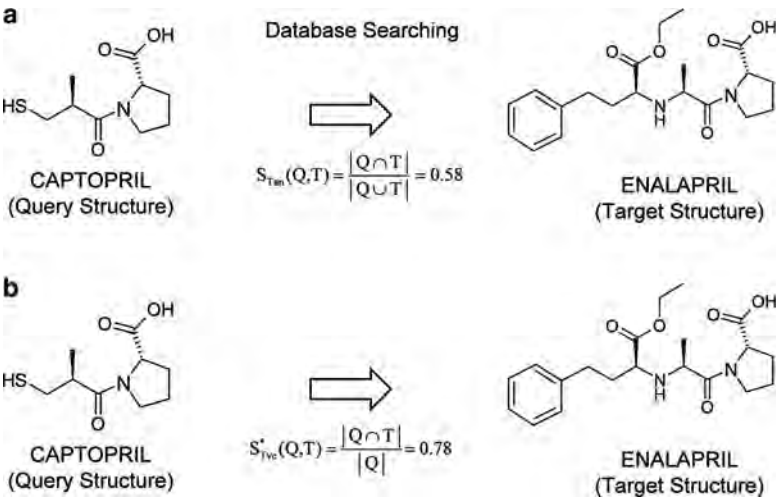


Fig. 3. Asymmetric similarity-based searching might provide some benefits not afforded by symmetric similarity-based searching. (**a**) Database searching using ISIS keys and symmetric (Tanimoto) similarity will not yield enalapril as a "database hit" because the similarity value (0.58) is too low. (**b**) In contrast, database searching using ISIS keys and asymmetric (Tversky) similarity could yield enalapril as a "database hit" because the asymmetric similarity value (0.78) is considerably larger than the corresponding symmetric one (0.58). This illustrates that small query molecules are more likely to retrieve larger target molecules in similarity searches based upon asymmetric rather than symmetric similarity indices.

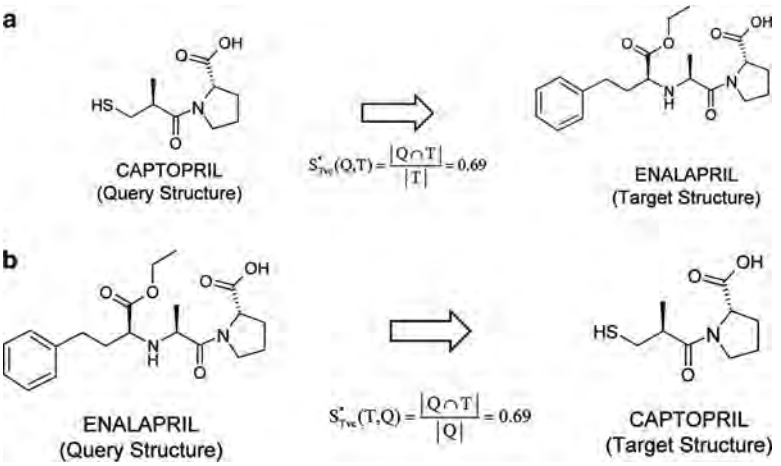


Fig. 4. (**a**) The other asymmetric (Tversky) similarity has a value of 0.69. Exchanging the roles of the query (Q) and target (T) molecules, i.e., Q(captopril) ⇒ T(enalapril) and T(enalapril) ⇒ Q(captopril), gives (**b**), which shows that large query molecules are more likely to retrieve smaller target molecules in similarity searches based upon asymmetric rather than symmetric similarity since the values of the corresponding indices are 0.69 and 0.58, respectively.

As is the case for asymmetric similarity indices, both $S_{\mathrm{Pet}_{\max}}(A,B)$ and $S_{\mathrm{Pet}_{\min}}(A,B)$are bounded by zero and unity, but are ordered with respect to each other and with respect to Tanimoto similarity, that is

$$0 \leqslant S_{\mathrm{Pet}_{\max}}(A,B) \leqslant S_{\mathrm{Tan}}(A,B) \leqslant S_{\mathrm{Pet}_{\min}}(A,B) \leqslant 1. \qquad (37)$$

2.2.3. Integer- and Categorical-Valued Feature Vectors

Feature vectors with integer- or categorical-valued components are identical in form to binary-valued vectors (*see* Eq. (18)). In contrast, however, each component takes on a finite number of values

$$v(x_k) = \begin{cases} \text{Finite, Ordered Set of Non - Negative Integers} \\ \text{Finite, Ordered Set of Values} \end{cases} \qquad (38)$$

In the integer case, these values usually refer to the frequency of occurrence of a given feature such as, for example, a molecular fragment. In the categorical case the values may refer to a binned variable. In both cases, the vectors live in discrete, lattice-like "hyper-rectangular" spaces, which are generalizations of the hypercubic spaces inhabited by bit vectors. Such spaces can also be described by multisets [50], but this formalism will not be used in this work.

Ideally, distances in these spaces should be based upon an $\ell_1$or city-block metric (*see* Eq. (20)) and not the ℓ_2 or Euclidean metric typically used in many applications. The reason for this are the

same as those discussed in Subsection 2.2.1 for binary vectors. Set-based similarity measures can be adapted from those based on bit vectors using a formula borrowed from fuzzy set theory [51, 52]. For example, the Tanimoto similarity coefficient becomes

$$S_{\mathrm{Tan}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) = \frac{\sum_k \min[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]}{\sum_k \max[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]} \tag{39}$$

As noted in Klir and Yuan [51] there are many possible denominators that can be used in place of $|A \cup B|$, each of which gives rise to a different similarity measure.

The asymmetric similarity coefficients become, in an analogous fashion [see Eqs. (30)]

$$\begin{aligned} S^{*}_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) &= \frac{\sum_k \min[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]}{\sum_k v_{\mathrm{A}}(x_k)} \\ S^{*}_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{B}},\mathbf{v}_{\mathrm{A}}) &= \frac{\sum_k \min[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]}{\sum_k v_{\mathrm{B}}(x_k)} \end{aligned} \tag{40}$$

As was the case in the previous section for bit vectors, it can be shown that the similarity coefficients defined here are also bounded,

$$0 \leqslant S_{\mathrm{Tan}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) \leqslant S_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}),\ S_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{B}},\mathbf{v}_{\mathrm{A}}) \leqslant 1. \tag{41}$$

in the case of non-negative integer-valued vector components. Other modifications are needed to accommodate non-integer values. Maggiora et al. [53], have discussed this issue, in general, for the case of field-based continuous functions, but their work also applies to "vectors" such as those described here.

In a methodology call holographic QSAR [54], integer-valued vectors are employed to characterize the frequency of occurrence of molecular fragments. The vectors are not, however, used in their "native" form but rather are folded into a smaller vector by hashing. Schneider et al. [55] have also used integer-valued vectors to characterize what they call 2-D pharmacophores.

Integer- and categorical-valued vectors can be converted into equivalent binary vectors by augmenting the components of a typical bit vector as shown in Fig. 5. The process is straightforward for integer-valued variables. Bajorath and co-workers [56] have developed a novel binning approach for variables with continuous values, basically converting them into categorical variables. Once the mapping to the augmented bit vector has been completed all of the usual bit-vector-based similarity measures (*see* Subsection 2.2.2 for further discussion) can be applied.

There are many other expressions for similarity that can be used for integer- and categorical-valued vectors. Again, the

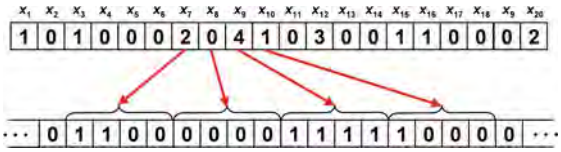


Fig. 5. In the scheme shown above, a 20-bit integer-valued vector (maximum integer value for each bit is 4) is converted into a 80-bit binary vector by converting each integer bit into a binary bit of 4-bit length. {0 = 0000; 1 = 1000; 2 = 1100; 3 = 1110; 4 = 1111}.

comprehensive discussion provided by Willett et al should be consulted for additional details [8]. Many of the features of discrete vector-based representations do not capture all of the relevant 3-D information in any substantive way, although they do capture some 3-D information indirectly, and this is why some feature vector procedures are referred to as "2.5-D" methods.

2.3. Continuous-Valued Feature Vectors

Vectors whose components have continuous values correspond to the more "traditional" types of vectors found in the physical sciences. They are of identical form to the discrete-valued vectors (*see* Eq. (18)) except that the components, $v_A(x_k)$, are continuous valued. In cheminformatics, however, the nature of the components is considerably different from those typically found in physics. For example, physiochemical properties, such as logP, solubility, melting point, molecular volume, Hammett $\sigma\rho$ parameters, surface charge, etc., as well as other descriptors derived explicitly for the purpose, such as BCUTs [35], have been routinely used. The use of continuous-valued vectors is usually confined to relatively low-dimensional chemical spaces, generally of less than ten dimensions (*see* Subsection 3 for further discussion). This is in sharp contrast to those discussed in the previous sections, where the dimensions are generally considerably larger.

Although it is ubiquitous in cheminformatics applications, the term vector should be used with caution as vectors are properly the objects of vector or affine spaces, and hence, must satisfy the axioms of these spaces. For example, vectors in BCUT chemical spaces do not form a vector space since the sum of two BCUT vectors may not lie in the space [35]. However, as long as this rather fine distinction is borne in mind significant problems should not arise, and the term vector, taken in its broadest if not strictest mathematical sense, will be used here. For a more general but brief discussion of vectors see the presentation of Euclidean vectors in Wikipedia [57].

2.3.1. Property-Based Continuous-Valued Feature Vectors

The components of most continuous-valued feature vectors are based on a variety of molecular properties such as solubilities, logPs, melting points, polar surface areas, molecular volumes,

various shape indices, and BCUTs, which are related to the charge, polarizability, and hydrogen bonding properties of molecules. Since these properties have a wide range of values they are typically scaled using the usual "z-transform" $z_i = (x_i - \bar{x})/\sigma_x$ favored by statisticians, where $\bar{x}$ is the average property-value and σ_x^2 is its variance (*N.B.* that this transformation is not strictly appropriate for multi-modal data). Other transforms have also been used; one of the most popular is $x'_i = (x_i - x_{\min})/(x_{\max} - x_{\min})$., where the values of the property, x_i, are mapped into the unit interval [0,1]. Simple scaling can be used to expand or contract the unit interval if desired.

An advantage of the z-transform is that it establishes a well-defined point of reference for the property-based vectors (the mean) as well as scaling the values of all of the variables to unit variance. BCUTs have a more complicated scaling, and the paper by Pearlman and Smith [35] should be consulted for further details. Since distances between vectors are invariant to the origin of the coordinate system, mean centering does not affect the result. However, the transformations used in all of the above procedures involve some form of scaling, and thus distances are not preserved between the original and scaled coordinate systems. Care must be exercised in the case of cosine similarity indices between vectors since they are both origin and scale dependent.

2.3.2. Inherent Non-orthogonality of Descriptor Coordinate Systems

An often-overlooked issue is the *inherent non-orthogonality* of coordinate systems used to portray data points. Almost universally a Euclidean coordinate system is used. This assumes that the original *variables* are orthogonal, that is are uncorrelated, when it is well known that this is generally not the case. Typically, PCA [12] is performed to generate a putative orthogonal coordinate system each of whose axes correspond to directions of maximum variance in the transformed space. This, however, is not quite correct. Since an orthogonal similarity transformation is used to carry out the PCA, and since such transformations rigidly rotate the original coordinate system, the angles among the coordinate vectors are unchanged. By exactly reversing the rigid rotation of the orthogonal principle-component coordinate system ones regenerates the original coordinate system, which is thus seen to be orthogonal. This clearly contradicts the general observation that most variables used in practice tend to be statistically correlated, that is are non-orthogonal. Importantly, even when the variables are properly uncorrelated this does not mean that they are necessarily *statistically independent* [58]. To correctly handle such correlated variables one must first orthogonalize the original variables, and then perform PCA to orient the orthogonal coordinate system along directions of maximum variance of the data points. This is rarely done in current practice, but what are the consequences of not doing this? As is well known from the theory

of tensors [59] both distances and angles between data vectors are affected by the angles between the coordinate axes (*Cf.* the discussion presented in Subsection 3.1.3 and in the paper by Raghavendra and Maggiora [95]). Conclusions drawn using, for example, either cosine similarity indices or distances will be affected *quantitatively* but not *qualitatively*. This is a manifestation of the fact that the topology (i.e. neighborhood relationships) of the space is preserved but its geometry (i.e. distances and angles) is not. The consequences of this are the following. The order of nearest-neighbors from a given reference molecule in a chemical space (*see* Subsection 3 for further details) will remain unchanged but the magnitude of their distances from the reference molecule will change. Thus, if one is only interested in, say, obtaining the 50 most similar molecules to a given reference molecule nothing will change by modifying the angles of the coordinate axes. If, one the other hand, one is interested in finding all molecules with similarities greater than or equal to, say, 0.85 with respect to that reference molecule, the results obtained will change since they depend on the angles of the coordinate vectors.

In many cases, however, problems brought about by skewed coordinate axes due to significant correlations among the variables are somewhat ameliorated by procedures, such as genetic algorithms, used for variable selection. While such procedures tend to remove highly correlated variables this may not always be the case so that coordinate system skew can still be a problem. However, if the variables are not too correlated the skew of a coordinate system will not significantly influence the overall results. A methodology is described in Subsection 3.3.1 that is based on molecular similarities and includes coordinate system non-orthogonality in a natural way.

2.3.3. Proximity Measures for Continuous-Valued Vectors

Because of the continuous nature of the vector components described in this section, other types of distance and similarity measures have been used. While the Hamming distance (*see* Eq. (20)) also applies for continuous vectors, Euclidean distances are usually used

$$\begin{aligned} d_{\text{Euc}}(\mathbf{v}_\text{A}, \mathbf{v}_\text{B}) &= \| \mathbf{v}_\text{A} - \mathbf{v}_\text{B} \| \\ &= \sqrt{\langle (\mathbf{v}_\text{A} - \mathbf{v}_\text{B}), (\mathbf{v}_\text{A} - \mathbf{v}_\text{B}) \rangle} \\ &= \sqrt{\sum_{k=1}^{n} \left(v_\text{A}(x_k) - v_\text{B}(x_k) \right)^2} \end{aligned} \tag{42}$$

In some instances, however, Minkowski distances are employed

$$d_{\text{Minkow}}(\mathbf{v}_\text{A}, \mathbf{v}_\text{B}) = \| \mathbf{v}_\text{A} - \mathbf{v}_\text{B} \|_{\ell_r} = \left[\sum_{k=1}^{n} \left| v_\text{A}(x_k) - v_\text{B}(x_k) \right|^r \right]^{\frac{1}{r}}, \tag{43}$$

where $r \geqslant 0$. Minkowski distances include both Hamming ($r = 1$) and Euclidean ($r = 2$) distances as special cases. Continuous distances can be converted into similarities using an appropriate monotonically decreasing function of distance, d, such as $\exp(-\eta \cdot d)$ or $1/(1 + \eta \cdot d)$, which both map to the unit interval, [0,1] for finite, non-negative values of η.

The most prevalent among the similarity coefficients is the so-called *cosine similarity index* or *correlation coefficient*. For the field-functions discussed in Subsection 2.4 it is usually called the *Carbó similarity index*

$$S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\sqrt{\|\mathbf{v}_A\|^2 \cdot \|\mathbf{v}_B\|^2}}, \qquad = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\|\mathbf{v}_A\| \cdot \|\mathbf{v}_B\|} \tag{44}$$

where the term in brackets in the numerator is the *inner product* of the two vectors

$$\langle \mathbf{v}_A, \mathbf{v}_B \rangle = \sum_{k=1}^{n} v_A(x_k) \cdot v_B(x_k) = \|\mathbf{v}_A\| \cdot \|\mathbf{v}_B\| \cos(\mathbf{v}_A, \mathbf{v}_B) \tag{45}$$

and their magnitudes are given by the Euclidean norm

$$\|\mathbf{v}_X\| = \sqrt{\langle \mathbf{v}_X, \mathbf{v}_X \rangle} = \sqrt{\sum_{k=1}^{n} v_X(x_k)^2}, \quad X = A, B. \tag{46}$$

It is important to note that the expressions in the latter two equations implicitly assume that the basis set used to describe the vectors is orthonormal.

As the similarity index is origin dependent there typically is a difference between the values computed for the cosine similarity index and correlation coefficients, since the latter is always computed at the mean of the of the data. Moreover, if the components of the vectors are all non-negative then $S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B)$ is also non-negative. When this is not the case, $S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B)$ may become negative, a situation that also obtains for the other similarity indices discussed in the remainder of this section. Maggiora et al. [53] have treated this case in great detail for continuous field functions, but the arguments can be carried through for finite vectors as well.

As has been pointed out numerous times, if $\mathbf{v}_A = \kappa \mathbf{v}_B$, then $S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B) = 1$ for all κ. This prompted Hodgkin and Richards [60] to define a slightly modified form of molecular similarity, usually called the Hodgkin similarity index, that does not suffer from this problem, namely,

$$S_{\mathrm{Hod}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\frac{1}{2}\left(\|\mathbf{v}_A\|^2 + \|\mathbf{v}_B\|^2\right)}. \tag{47}$$

Petke [41] has developed two additional indices that bound both the Carbó and Hodgkin similarity indices, namely

$$S_{\mathrm{Pet}_{\min}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\min\left(\|\mathbf{v}_A\|^2, \|\mathbf{v}_B\|^2\right)} \tag{48}$$

and

$$S_{\mathrm{Pet}_{\max}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\max\left(\|\mathbf{v}_A\|^2, \|\mathbf{v}_B\|^2\right)}, \tag{49}$$

that are analogous to those given, respectively, in Eqs. (28) and (29) for the case of sets or binary vectors. Recently, a comprehensive analysis has been given for continuous, field-based functions of all of the similarity coefficients of this general form, which characterizes their linear ordering and their upper and lower bounds [53] (*see* Eq. (66)). Their approach can be taken over in its entirety to the case of finite-dimensional vectors covered in this section. Thus, the bounds of the similarity indices in Eqs. (44), (47), (48), and (49), are given by

$$0 \leqslant S_{\mathrm{Pet}_{\max}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant S_{\mathrm{Hod}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant S_{\mathrm{Pet}_{\min}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant \infty. \tag{50}$$

All of the indices except $S_{\mathrm{Pet}_{\min}}(\mathbf{v}_A, \mathbf{v}_B)$ have upper bound of unity.

2.4. Field-Based Functions

Many methods exist for assessing 3-D molecular similarity. Good and Richards [61] and Lemmen and Lengauer [62] provide comprehensive reviews of most of the methods in use today, a large class of which utilizes some form of vector-based representation of 3-D molecular features such as 3-D pharmacophores [33, 63] and various types of 3-D shape descriptors [64]. The components of these vectors can be binary, integer, categorical, or continuous as discussed in the previous sections. Most 3-D methods, however, involve some type of direct alignment of the molecules being considered. Early on RMS deviations between specific atoms in the molecules being compared were employed, but this required identifying the key atoms, a non-trivial computational task. A variety of other 3-D methods exist [62], but the bulk of the 3-D methods utilize some form of field-based function to represent the fields or *pseudo*-fields surrounding the molecules that can be either continuous or discrete. Examples include "steric," electrostatic potential and "lipophilic" fields [39]. A novel, albeit lower resolution approach, based on ellipsoidal Gaussians has recently been developed [65] and shows great promise as a means for handling very large sets of molecules. Several workers have also developed a field-based methodology for directly aligning molecules based

upon their electric fields [41, 60] that differs from the usual scalar potential fields that are typically matched.

Interestingly, there is a close analogy between the alignment of 3-D molecular fields and the determination of maximum common substructures of two chemical graphs (*see* Subsection 2.1). Both cases involve the search for optimal overlays or alignments. The former requires continuous optimizations of non-linear similarity indices that give rise to large numbers of solutions and to great difficulties in clearly identifying the global maximum or "best" solution (*see* Subsection 2.4.3). The latter requires discrete optimizations, but the problem is NP complete and thus does not scale well computationally.

A major factor differentiating 3-D from 2-D similarity methods, regardless of the type of 3-D method employed is the need to account in some manner for conformational flexibility. There are two ways this is generally accomplished. One method involves carrying out a conformational analysis and selecting a subset of "appropriate" conformations for each molecule. All pairwise alignments of the selected conformations are then computed [37]. The other method involves some form of conformational search carried out simultaneously with the alignment process [66, 67]. Because of its importance in similarity-based alignments of molecules the remainder of the discussion in this section will focus on field-based methods.

2.4.1. Representation of Molecular Fields

Field-based methods generally utilize linear combinations of appropriate functions that are associated in some way with the atoms of the molecule under study:

$$F_{\mathrm{A}}^{\alpha}(\mathbf{r}) = \sum_{i \in \text{atoms}} a_i^{\alpha} f_i(\mathbf{r}), \tag{51}$$

where "α" designates the specific type of field or property being considered. The coefficients a_i^{α} weight the atom-based functions and in many cases are used to characterize specific properties attributed to the individual atoms (*vide infra*). Unnormalized, spherically symmetric Gaussian functions, "Gaussians" for short, are by far the most ubiquitous functions used in field-based applications:

$$f_i(\mathbf{r}) = \exp\left(-\kappa_i |\mathbf{r} - \mathbf{R}_i|^2\right), \tag{52}$$

where $\mathbf{R}_i$ is the location of the Gaussian, generally at an atomic center, and κ_i is its "width," which is the reciprocal of the variance, that is $\kappa_i = 1/\sigma_i^2$. The variance is sometimes referred the orbital radius, $\rho_i = \sigma_i^2$, of a Gaussian [68]. As $\kappa_i \to 0$, $f_i(\mathbf{r})$ becomes more spread out and conversely as $\kappa_i \to \infty$, $f_i(\mathbf{r})$ becomes sharper until, in the limit, it approaches an infinitely sharp delta function. In the latter case, atoms are essentially represented as points, while

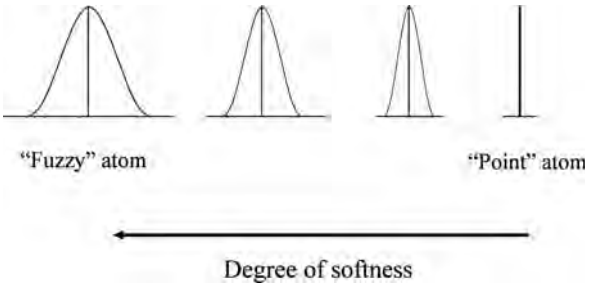


Fig. 6. Gaussian curves as a function of increasing width. As the degree of softness increases the curves represent "fuzzy" atoms and as the degree of softness deceases the Gaussian converges to a "point" atom model.

in the former case, they are represented as "soft spheres" as illustrated in Fig. 6. Although applications utilizing ellipsoidal Gaussians, which represent a generalization of spherically-symmetric Gaussians, are finding increasing application in chem-informatics [65], the focus of the current chapter will remain on the former more ubiquitous and simpler functions.

A useful property of Gaussians is that the integral of the product of two Gaussians [69] is given by another Gaussian that is a function of their distance of separation

$$\int f_i(\mathbf{r}) \cdot f_j(\mathbf{r}) d^3\mathbf{r} = \left(\frac{\pi}{\kappa_i + \kappa_j}\right)^{3/2} \exp\left(-\frac{\kappa_i \kappa_j}{\kappa_i + \kappa_j} \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right). \tag{53}$$

Thus, the "overlap" of two molecules, A and B, with respect to the field of property α, $\Omega(F_A^\alpha, F_B^\alpha)$, is given by

$$\begin{aligned} \Omega(F_A^\alpha, F_B^\alpha) &= \int F_A^\alpha(\mathbf{r}) \cdot F_B^\alpha(\mathbf{r})\, d^3\mathbf{r} \\ &= \sum_{i \in A} \sum_{j \in B} a_i^\alpha \cdot b_j^\alpha \int f_i(\mathbf{r}) f_j(\mathbf{r})\, d^3\mathbf{r} \\ &= \sum_{i \in A} \sum_{j \in B} a_i^\alpha \cdot b_j^\alpha \left(\frac{\pi}{\kappa_i + \kappa_j}\right)^{3/2} \exp\left(-\frac{\kappa_i \kappa_j}{\kappa_i + \kappa_j} \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right) \end{aligned}, \tag{54}$$

$$\Omega(F_A^\alpha, F_B^\alpha) = \sum_{i \in A} \sum_{j \in B} \tilde{a}_i^\alpha \cdot \tilde{b}_j^\alpha \exp\left(-\frac{\kappa_i \kappa_j}{\kappa_i + \kappa_j} \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right), \tag{55}$$

where the modified coefficients, $\tilde{a}_i^\alpha$ and $\tilde{b}_j^\alpha$, are obtained by including the square root term equally into the two field (i.e., property) coefficients a_i^α and b_j^α given in Eq. (54). In most cases the width parameters, κ_i and κ_j, are chosen to be the same for all atoms.

Equation (55) is a general form that is used in a number of field-based approaches to 3-D molecular alignment and similarity.

For example, in the program Seal [70] the coefficients given either in Eq. (54) or Eq. (55) are subsumed into a single "property coefficient," $\tilde{a}_i^{\alpha} \cdot \tilde{b}_j^{\alpha} \Rightarrow w_{i,j}$, which may account for the effect of multiple types of properties,

$$\Omega_{\text{'Seal'}}(\text{A}, \text{B}) = \sum_{i \in \text{A}} \sum_{j \in \text{B}} w_{i,j} \exp\left(-\kappa \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right). \tag{56}$$

The exponential coefficient, κ, determines the spread of the Gaussian and is taken be identical for all atom pairs. Some methods assign property values directly to the coefficients $\tilde{a}_i^{\alpha}$ and $\tilde{b}_j^{\alpha}$ [67, 71].

An alternative approach [37] treats the steric and electrostatic potential fields directly. The steric field is generally given by an expression similar to that in Eq. (51),

$$F_{\text{A}}^{\text{st}}(\mathbf{r}) = \sum_{i \in \text{atoms}} a_i^{\text{st}} f_i(\mathbf{r}), \tag{57}$$

where the coefficients are usually taken to be unity, that is $a_i^{\text{st}} = 1$, the field functions, $f_i(\mathbf{r})$, are usually taken to be Gaussians (*see* Eq. (52)), and the width parameters, κ_i, are either held constant for all atoms or are adjusted for each specific "atomic environment" [37]. In the case of the molecular electrostatic potential ("el") field

$$F_{\text{A}}^{\text{el}}(\mathbf{r}) = \sum_{i \in \text{atoms}} \frac{q_i}{|\mathbf{r} - \mathbf{R}_i|} \tag{58}$$

the $1/r$ term, which becomes singular at each atomic nucleus, presents a computational problem that was solved by Good et al. [72], who developed a Gaussian expansion of the "$1/r$" term,

$$\frac{1}{|\mathbf{r} - \mathbf{R}_i|} \approx \sum_{k \in \text{A}_i} c_k f_{i,k}(\mathbf{r}), \tag{59}$$

that significantly expedites computations. In this expression $f_{i,k}(\mathbf{r})$ is the k-th Gaussian in the expansion of $1/r$ about the i-th atom, A_i, of molecule A. The expansion usually consists of two or three terms, and the expansion coefficients, c_k, are obtained by least-squares minimization. Note that the width parameters, κ_k, are independent of the atom center and differ significantly from each other in order to fit the $1/r$ term with sufficient accuracy [72]. Substituting Eq. (59) into Eq. (58) converts it into a sum of Gaussians, and thus most field-based similarity measures (*vide infra*) only require calculation of Gaussian overlap integrals (*see* e.g., Eqs. (53) and (54)) when dealing with steric or electrostatic potential fields. Thus, many of the issues that plague similarity calculations carried out within a discrete lattice framework are no longer a problem in the case of continuous field-based functions.

2.4.2. Field-Based Similarity Indices

Field-based similarities are usually evaluated by the cosine or correlation function similarity measure employed initially by Carbó and Calabuig [73] to compute molecular similarities based upon quantum mechanical wavefunctions. Such a measure, which is usually called a Carbó similarity index, is given by

$$S_{\mathrm{Car}}\left(F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right\rangle}{\left\|F_{\mathrm{A}}^{\alpha}\right\| \cdot \left\|F_{\mathrm{B}}^{\alpha}\right\|} = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right\rangle}{\sqrt{\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2} \cdot \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}}}, \tag{60}$$

where the inner product in the numerator is now given by an integral rather than a summation since the objects considered here are field functions, F_{A}^{α} and F_{B}^{α}, not vectors, although strictly speaking functions are equivalent to infinite dimensional vectors,

$$\left\langle F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right\rangle = \int F_{\mathrm{A}}^{\alpha}(\mathbf{r}) \cdot F_{\mathrm{B}}^{\alpha}(\mathbf{r}) d^{3}\mathbf{r}, \tag{61}$$

and the Euclidean norm of the functions is given by

$$\left\|F_{X}^{\alpha}\right\| = \sqrt{\int F_{X}^{\alpha}(\mathbf{r})^{2} d^{3}\mathbf{r}}, \quad X{=}A, B. \tag{62}$$

Note the similarity of Eqs. (45) and (46) to Eqs. (61) and (62). The main difference between them arises in the way in which the inner products are evaluated. Also, as was the case for vectors if the field functions are non-negative functions $S_{\mathrm{Car}}(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha})$ will be non-negative. When this is not the case, however, $S_{\mathrm{Car}}(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha})$ may become negative, a situation that also obtains for the other similarity indices discussed in the remainder of this section. Maggiora et al. [53], have treated this case in great detail for continuous field functions, but the arguments can be carried through for finite vectors as well (*vide supra*).

As discussed in the previous section for vectors, if F_{A}^{α} and F_{B}^{α} differ only by a constant, that is if $F_{\mathrm{A}}^{\alpha} = K \cdot F_{\mathrm{B}}^{\alpha}$, then $S_{\mathrm{Car}}(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}) = 1$ regardless of the specific form of the functions. While this is not a likely occurrence in practical applications, Hodgkin and Richards [60] nonetheless defined a slightly altered similarity measure, usually referred to as the Hodgkin similarity index, which is not affected by this problem and is given by

$$S_{\mathrm{Hod}}\left(F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right\rangle}{\frac{1}{2}\left(\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2} + \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}\right)}, \tag{63}$$

where the terms in the denominator of Eqs. (60) and (63) are, respectively, the geometric and arithmetic means of the squared norms of F_{A}^{α} and F_{B}^{α}. As has been shown by Maggiora et al. [53],

a family of similarity indices can be defined in terms of the means of the squared norms in their denominators.

The Petke indices, defined earlier for vectors (*see* Eqs. (48) and (49)), are given by

$$S_{\mathrm{Pet}_{\min}}\left(F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right\rangle}{\min\left(\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2}, \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}\right)} \tag{64}$$

and

$$\mathrm{S}_{\mathrm{Pet}_{\max}}\left(F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right\rangle}{\max\left(\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2},\ \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}\right)}. \tag{65}$$

All of the same bounding properties described in the previous section for vectors (*see* Eq. (50)) obtain here as well, including the fact that all of the indices except $S_{\mathrm{Pet}_{\min}}\left(F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right)$ are bounded from above by unity [53]:

$$0 \leqslant S_{\mathrm{Pet}_{\max}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant S_{\mathrm{Hod}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant S_{\mathrm{Car}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant S_{\mathrm{Pet}_{\min}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant \infty. \tag{66}$$

None of the cosine/correlation-like similarity indices or their complements (*see* Subsection 2.5 on Dissimilarity Measures) are true metrics; that is, they do not obey the distance axioms. Petitjean [74, 75], however, has developed a distance-based methodology, but it has not been applied in many cases.

Similarity indices corresponding to different fields can be combined into an overall similarity index, for example

$$S(F_{\mathrm{A}}, F_{\mathrm{B}}) = \lambda \cdot S(F_{\mathrm{A}}^{\mathrm{st}}, F_{\mathrm{B}}^{\mathrm{st}}) + (1 - \lambda) \cdot S(F_{\mathrm{A}}^{\mathrm{el}}, F_{\mathrm{B}}^{\mathrm{el}}), \tag{67}$$

where S is the Carbó, Hodgkin, Petke, or other appropriate index [53] and λ is the weighting coefficient. Mestres et al. [37], have used a value ($\lambda \approx 0.66$), arrived at pragmatically, that weights steric to electrostatic-potential similarity in a 2:1 ratio.

2.4.3. Deriving Consistent Multi-molecule Alignments

As has been shown by Mestres et al. [29], the optimal solution, $S_X(F_{\mathrm{A}}, F_{\mathrm{B}})_1$, may not correspond to the correct "experimentally-derived" molecular alignment. To address this problem, these authors developed the concept of *pairwise consistency*, which is depicted in Fig. 7. Consider the similarities of three molecules A, B, and C. Suppose molecule A is the reference molecule, which is held fixed, and molecules B and C are the adapting molecules. Now determine the optimal similarity solutions for $S(F_{\mathrm{A}}, F_{\mathrm{B}})_1$ and $S(F_{\mathrm{A}}, F_{\mathrm{C}})_1$ using an appropriate similarity index. Both molecules B and C are now aligned to molecule A. Keeping their positions relative to molecule A fixed, compute $S(F_{\mathrm{B}}, F_{\mathrm{C}})^{*}$, which is not necessarily equal to the optimized solution, that is $S(F_{\mathrm{B}}, F_{\mathrm{C}})^{*} \neq S(F_{\mathrm{B}}, F_{\mathrm{C}})_1$. Pairwise consistency holds only in the

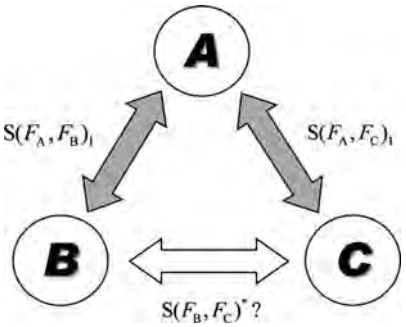


Fig. 7. Depiction of *pairwise consistency* among three molecules, A, B and C.

case when *equality* obtains, otherwise the solutions are said to be pairwise inconsistent. Sometimes pairwise consistency is obtained when one of the lower similarity solutions is considered, say for example $S(F_A, F_C)_2$. In such cases, the alignments given by $S(F_A, F_B)_1$,$S(F_A, F_B)_2$, and $S(F_B, F_C)^* = S(F_B, F_C)_1$ are assumed to be the correct alignments. In many cases, it is not possible to identify pairwise consistent set of solutions. In such cases, the fields of three molecules are *simultaneously aligned* using the average of the pairwise similarities

$$S(F_A, F_B, F_C) = \tfrac{1}{3}[S(F_A, F_B) + S(F_A, F_C) + S(F_B, F_C)]. \qquad (68)$$

This procedure automatically generates pairwise consistent solutions, and it can be continued to higher orders until consistent solutions are obtained to all orders, a computationally very demanding task that has not been pursued in most cases since the ternary similarities are generally sufficient for molecular design purposes.

Note that any 3-D similarity method that involves molecular superpositioning can be used to assess the consistency of multi-molecule alignments. However, 3-D methods that do not involve superpositions (e.g., see the interesting recent work by Ballester and Richards [76]) are not suitable for this type of analysis.

2.4.4. Addressing Conformational Flexibility

As noted in the introduction of this section, computing 3-D molecular similarities of a set of molecules requires physically aligning the appropriate fields of the molecules, while accounting for their conformational flexibility, which can be accomplished in two ways either by rigid body superpositions of selected conformations of each of the molecules being aligned or by simultaneous conformational sampling during the alignment process. In the rigid-body case, one molecule is generally chosen as the *Reference Molecule*, which remains fixed, while the other *Adapting Molecule* is translated and rotated until a maximum of the similarity index is obtained. Since the similarity index is a non-linear function it generally has multiple solutions,

$$S(F_A, F_B)_1 \geqslant S(F_A, F_B)_2 \geqslant \cdots \geqslant S(F_A, F_B)_k \geqslant \cdots \tag{69}$$

although it is difficult to know if the global maximum has been attained. To increase the chances that all of the best solutions are obtained, multiple starting geometries are usually sampled.

An added difficulty in rigid-body alignment is that all "relevant" conformations, N, of the reference molecule must be aligned with all "relevant" conformations, M, of the adapting molecule – $N \times M$ alignments must be carried out where, as discussed above, each alignment involves multiple starting geometries. This is a significant computational burden for the alignment of a single pair of molecules, and thus carrying out alignments for a large set of molecules is not computationally feasible at this time.

There are some approaches that hold promise for speeding up the computations. A novel procedure based upon Fourier transforms was developed by Nissink et al. [77], and used by Lemmen et al. [71] The method separates the translational and rotational motions needed to align pairs of molecules and thus allows the separate optimization of each, thereby facilitating the overall alignment process. While this certainly speeds up the computations, it does not significantly alter the significant time requirements of rigid body alignments.

An alternative approach that combines conformational searching with similarity-based structure alignment perhaps holds more promise in terms of speeding up the process of aligning conformationally flexible molecules. In contrast to the rigid-body alignment process described above, in this case both molecules are treated on an equal footing and are allowed to move and conformationally flex. In the approach of Blinn et al. [66], which is similar to that developed by Labute [67], the energy of the combined system of the two molecules being aligned, $E_{A,B}^{total}$, is given by

$$E_{A,B}^{total} = E_A^{conf} + E_B^{conf} + E_{A,B}^{sim}, \tag{70}$$

where E_A^{conf} is the conformational energy of molecule A, E_B^{conf} is the conformational energy of molecule B, and $E_{A,B}^{sim}$ is a pseudo-energy penalty term, which is given by

$$\begin{aligned} E_{A,B}^{sim} &= K_{sim} \cdot [1 - S(F_A, F_B)] \\ &= K_{sim} \cdot D(F_A, F_B) \end{aligned}, \tag{71}$$

where K_{sim} is an adjustable proportionality constant, which lies in the range of 5–20 kcal/mol. The dissimilarity $D(F_A, F_B)$ (*see* Subsection 2.5) is used rather than similarity since the penalty term should vanish when the fields of the two molecules are in perfect alignment that is when $S(F_A, F_B) = 1 \rightarrow D(F_A, F_B) = 0 \rightarrow E_{A,B}^{sim} = 0$. Alternatively, the maximum penalty should be assessed when

$$S(F_A, F_B) = 0 \rightarrow D(F_A, F_B) = 1 \rightarrow E_{A,B}^{sim} = K_{sim}. \tag{72}$$

Other forms for the pseudo-energy penalty term have also been investigated [66, 67]. In any case, the pseudo-energy penalty term acts as a constraint on the overall energy of the system, which is a balance between favorable conformational energies and overall molecular alignment as measured by field-based similarity (dissimilarity).

While the approaches noted above use some type of stochastic procedure to explore the similarity-constrained conformational space [66, 67], they are not in principle restricted to such search methods. Molecular dynamics-based procedures are also possible, although to our knowledge none have as yet been developed to address this problem.

2.4.5. Multi-conformer-Dependent Similarities

It may be argued that 3-D similarities most closely represent the concept of molecular similarity. Since many molecules can attain multiple conformational states, it is not unreasonable to assume that these low-lying conformational states should play a role in molecular similarity. To date, 3-D similarity methods typically choose a single conformer, usually the one with the lowest conformational energy. This section describes a possible approach to the question "Given that a set of conformer-dependent similarities can be determined, how does one assign an aggregate similarity to the set of values"? It is not meant to be a finished work, but rather is meant to encourage further work and discussion on what undoubtedly will become a more important issue as computational methods improve and become faster (*see* e.g. the work described in [76]).

One, but certainly not the only, approach is to weigh the similarities by the joint probability of the two conformers,

$$\langle S(F_{\mathrm{A}}, F_{\mathrm{B}})\rangle = \sum_{i=1}^{n_{\mathrm{A}}} \sum_{j=1}^{n_{\mathrm{B}}} \mathrm{Prob}\left(\mathrm{A}_i, \mathrm{B}_j\right) \cdot \mathrm{S}(F_{\mathrm{A}_i}, F_{\mathrm{B}_j}), \tag{73}$$

where A_i is the i-th conformational state of molecule A and B_j the j-th conformational state of molecule B. In the case where the conformations of each of the two molecules are determined independently, as described in Subsection 2.4.4,

$$\mathrm{Prob}\left(\mathrm{A}_i\,, \mathrm{B}_j\right) = \mathrm{Prob}(\mathrm{A}_i) \cdot \mathrm{Prob}\left(\mathrm{B}_j\right). \tag{74}$$

The two conformational-state probabilities can be estimated using a Boltzmann or other suitable type of probability function; the former is given by [78]

$$\mathrm{Prob}(\mathrm{M}_\ell) = \frac{\exp(-\kappa \cdot \Delta E_{\mathrm{M}_\ell})}{Z_{\mathrm{M}}}, \quad \mathrm{M} = \mathrm{A},\ \mathrm{B},\ \mathrm{C}, \ldots; \quad \ell = 1, 2, \ldots, n_{\mathrm{M}} \tag{75}$$

where κ is Boltzmann's constant, $\Delta E_{\mathrm{M}_\ell}$ is the energy (in kcal/mol) of the i-th conformation of M *relative* to its lowest energy

conformation, and Z_M is the conformational partition function for molecule M, i.e.

$$Z_M = \sum_{\ell=1}^{n_M} \exp(-\kappa \cdot \Delta E_{M_\ell}), \quad M = A, B, C, \ldots \tag{76}$$

where n_M is the number of conformational states considered for molecule M. Since n_M is usually less than the actual number of conformational states, Z_M is only an estimate of the true conformational partition function. However, this may not be the most serious limitation of the proposed approach as conformational energies may also be poorly estimated and solvent effects, when they are explicitly considered, also can represent a serious source of error (*vide infra*).

For entirely practical reasons of computational efficiency, conformational energies are usually computed in the gas phase using molecular mechanics potential-energy functions [79]. However, recent advances in computational methods have made solvent-based conformational energetics computationally feasible [80]. Methods have also been developed for computing Gibbs free energies [81], which are physically more realistic. However, even in cases where conformational energetics are computed with reasonable accuracy, the number of conformers considered is usually only a subset of the possible conformers such as those, for example, based on clustering (*see* discussion in Subsection 2.2.4). Thus, the calculated probabilities are highly approximate. Nevertheless, this approach accounts in some, albeit very approximate, fashion for the role that multiple conformations can play in MSA. Whether this more elaborate approach produces better results than those obtained using single low-energy conformations has yet to be investigated.

Considering the combinatoric explosion that can arise in the "independent conformer" approach, it may be more effective to compute conformationally-weighted similarities in a more direct manner. Such an approach may be developed from that described in Subsection 2.2.4 (*see* Eqs. (70)–(72)) [66] or some other variant [67] that combines conformational searching simultaneously with similarity-based alignment. Although, conformational independence (*vide supra*) is lost in this approach, it may not be an impediment to obtaining computationally-feasible, conformationally-weighted similarities.

A related but "softer" approach (*see* e.g., Petit et al. *submitted* [82], for a discussion of soft approaches to the Rule of Five that is relevant here) is to consider the conformationally-dependent distribution of similarity values between each pair of molecules under consideration in a given study, and then to compare the resulting similarity or cumulative similarity distributions [83] directly or with respect to several relevant distributional parameters such

as the mean, median, standard deviation, or other statistical moments [84].

2.5. Dissimilarity Measures

Dissimilarity is generally taken to be the complement of similarity, that is

$$D(\mathrm{A},\mathrm{B}) = 1 - S(\mathrm{A},\mathrm{B}). \tag{77}$$

While this is mathematically reasonable, and is thus used extensively, psychologically the two concepts are not so simply related. This stems from the fact that assessing the similarity of two objects is easier than assessing their dissimilarity. As two objects become less and less similar a point is reached, say for a similarity value of 0.35, below which it is very difficult to assign a value to their similarity. Since dissimilarity is just the complement of similarity it follows that humans can only properly assess the dissimilarity of two objects if they are not too dissimilar. Thus, even though we can assign dissimilarities ~1.0 using Eq. (77) its "meaning" is not easily grasped. This is why we have focused our discussion on similarity rather than dissimilarity, even though the concept of dissimilarity has important practical applications in studies of molecular diversity. Martin [85] has edited an interesting account of the development and implementation of the concepts of molecular dissimilarity and diversity in cheminformatics and combinatorial chemistry. Seilo [86] has also investigated some of the issues associated with comparing dissimilar molecules.

3. Chemical Spaces

This section on chemical spaces, while important, is not presented with the same level of mathematical detail as given in earlier sections. The object here is to provide a general discussion of some of the important characteristics of these spaces. A recent review by Medina-Franco et al. [87] presents an excellent overview of many aspects of chemical spaces relevant to drug design research; Oprea and Gottfries [88] published the first paper on navigating chemical spaces.

The concept of a chemical space derives from the notion of a space used in mathematics and is taken here to be a set of molecules along with one or more relationships defined on the elements (i.e. molecules) of the set. The nature of a given chemical space depends, directly or indirectly, on how the molecular information is represented (Subsection 2); the representation used strongly influences what can be known about the set of molecules under study. Figure 8 illustrates the dramatic effect that different molecular representations can have on the distribution of compounds in chemical space (*See* figure legend for details and

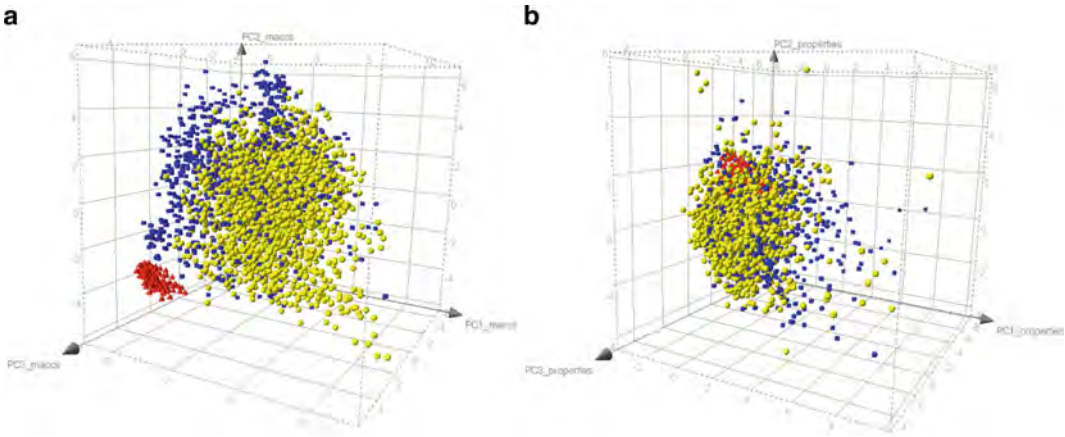


Fig. 8. Comparison of the 125 new molecules (*red triangles*), 1,490 approved drugs (*blue squares*) and a representative set of 2,000 diverse compounds (*yellow spheres*) from Molecular Libraries Small Molecules Repository (MLSMR) (*Original figure provided courtesy of Dr. José Medina-Franco*). (**a**) Depiction of a visual representation of the chemical space obtained by PCA of the similarity matrix computed using MACCS keys and Tanimoto similarity. The first three PCs account for 62.1% of the variance. (**b**) Depiction of a visual representation of the property space obtained by PCA of six scaled physicochemical properties (MW, RB, HBA, HBD, TPSA, and SlogP). The first three PCs account for 84.3% of the variance.

Subsections 3.1 and 3.1.1 for a discussion of how chemical spaces are depicted and how reduced-dimensional chemical spaces are constructed.). Panel (a) of the figure shows that the set of 125 new molecules (red triangles) is *structurally different* from the 1,490 approved drugs (blue cubes) [89] and the representative set of 2,000 compounds from the MLSMR [90]. Panel (b) indicates that the new library is located within a region of the physicochemical property-based chemical space that is associated with bioactive molecules. Both of these pieces of information are useful. Nevertheless, the significant differences in the distribution of compounds in these two spaces clearly show the crucial role that representation plays in defining chemical spaces.

Importantly, unlike the case in physics, the underlying relationships in chemical spaces are not invariant to representation. For example, neighborhood relationships that obtain in one chemical space may not also obtain in another chemical space [2] (*Cf.* Patterson et al. [91]). This is shown in Fig. 9, which schematically depicts the chemical spaces generated for the same set of molecules using two different representations (♣ and ♠). As is seen in the figure, molecules that are nearest-neighbors with respect to one representation may not even be close neighbors in the other (*See*, for example, mappings ❶ and ❷ and mappings ❷ and ❸ of Fig. 9). Thus, there is *loss of topological invariance*, which is a much more serious condition than the loss of purely geometric features such as the distances between molecules or the angles between vectors representing the locations of molecules in

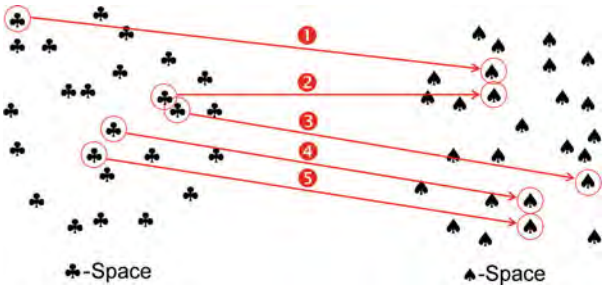


Fig. 9. An example illustrating that chemical spaces are representation dependent. Five point-to-point mappings are indicated on the figure. Mappings ❶ and ❷ show that two molecules that are well separated in ♣-Space may be nearest-neighbors in ♠-Space, while mappings ❷ and ❸ show that two molecules that are nearest-neighbors in ♣-Space are well separated in ♠-Space. Mappings ❹ and ❺ show that neighborhood relationships may also be maintained by mappings between the two spaces.

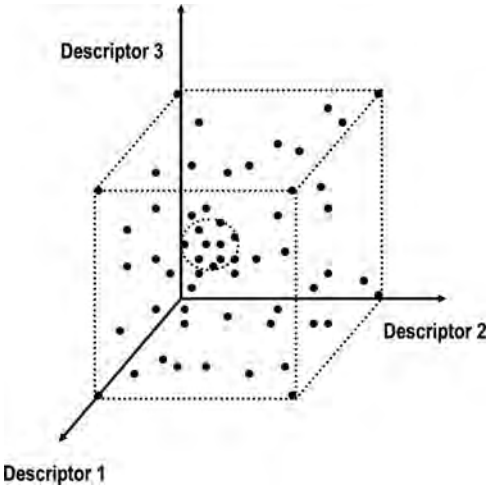


Fig. 10. In coordinate-based chemistry space, points in close proximity are considered to be similar. For instance, the compounds within the sphere shown here are quite similar to each other compared to compounds in the extremities of this 3-dimensional chemistry space.

chemical spaces. Loss of topological invariance can have dire consequences in subset selection procedures since it can change the rank ordering of neighboring compounds with respect to chemical spaces constructed using different similarity measures [2].

3.1. Dimensionality of Chemical Spaces

Chemical spaces can be grouped into two broad classes, namely, *coordinate-based* and *coordinate-free*. In coordinate-based chemical spaces molecules are represented as points distributed throughout the space as illustrated in Fig. 10. Points in close proximity are considered to represent similar molecules,

while distant points represent dissimilar molecules. An important feature of coordinate-based chemical spaces is that the *absolute position* of a molecule within the space is known, not just its position relative to the other molecules in the space. This is not the case with coordinate-free chemical spaces. In such spaces the relationship of a given molecule to its near and far neighbors is known but not its location within the space. Thus, finding "compound voids" in a coordinate-free chemical space is a much more difficult task than it is in a coordinate-based chemical space. An additional useful feature of coordinate-based chemical spaces is their ability to portray the distribution of compounds in ways that, in many cases, can enhance our understanding of the space. However, as is discussed in the following paragraph, the high-dimensionality of these spaces can frustrate attempts to visualize them.

The dimension of a coordinate-based chemical space is simply the number of independent variables used to define the space. As seen in earlier discussions, the dimension of such spaces can be quite large, and there are a significant number of examples where the dimension can exceed one million [33, 48]. Even for spaces of much lower dimension, say around ten or greater, the effects of the "*Curse of Dimensionality*" [92, 93] can be felt. Bishop [94] provides an excellent example, which shows that the ratio of the volume of a hypersphere inscribed in a unit hypercube of the same dimension goes to zero as the dimensionality goes to infinity. Actually, even for the ten-dimensional case the volume of the hypersphere is less than ten percent of that of the corresponding hypercube. Raghavendra and Maggiora [95] provide a more detailed discussion of some of the idiosyncratic behaviors of high-dimensional spaces [96]. In addition, most of the spaces of extremely high dimension are discrete since the number of "points" (i.e. molecules) in the space is finite, a feature that can present additional problems.

Although, it is possible in coordinate-free chemical spaces to rigorously construct coordinates within a Euclidean space for any set of molecules in the space, faithfully representing the intermolecular proximities may require that the space be of quite high-dimension, in an extreme case possibly equal to one less than the number of molecules in the set. As will be discussed in the following section, a number of methods exist for constructing low-dimensional Euclidean spaces for both high-dimensional coordinate-based and coordinate-free representations of molecules.

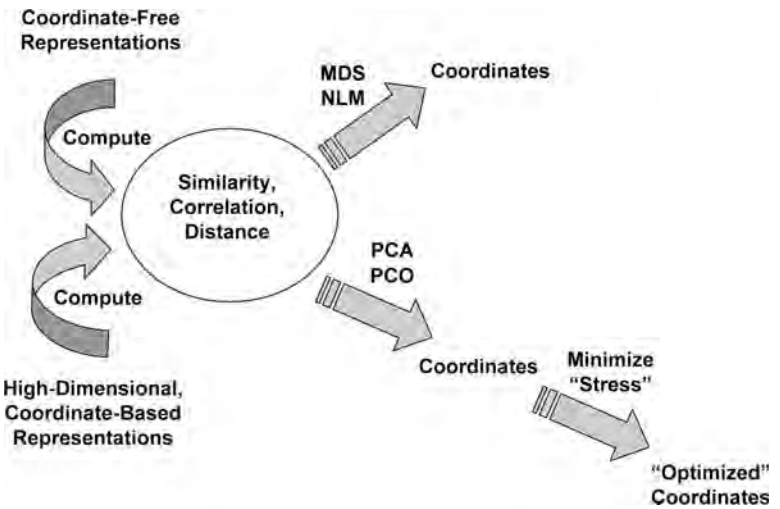
3.2. Constructing Reduced-Dimension Chemical Spaces [97]

This section provides a brief account on the construction of reduced-dimension chemical spaces for sets of molecules described by coordinate-free or by high-dimensional coordinate-based representations. Inherently low-dimensional chemical spaces such as those generated, for example, by BCUT descriptors

are not considered; the paper by Pearlman and Smith [35] should be consulted for a discussion of these descriptors. It is important to note that all of the issues surrounding the inherent non-orthogonality of coordinate systems described in Subsection 2.3.2 are applicable here as well, and that section should be consulted for further details.

Scheme 1 illustrates the various procedures for the construction of reduced-dimension coordinate systems – similarity (dissimilarity), correlation, or distance play a central role in these coordinate systems. The first step in reducing the dimension of either coordinate-free or high-dimensional coordinate-based representations is computation of some proximity measure of the similarity, correlation, or distance between all the pairs of molecules in the set of interest. This can be accomplished using the methods described earlier in this work. For example, in the coordinate-free case similarity can be computed using the graph-theoretical procedures described in Subsection 2.1, the field-based approach described in Subsection 2.4, or other less well-known approaches such as shape-group [98] and feature-tree [99] methods. In high-dimensional coordinate-based cases all of the vector-based approaches described in Subsection 2 are applicable (*N.B.* that field-based approaches can also fit under this rubric, since field-based functions can be considered to be infinite-dimensional vectors).

Once a proximity measure has been computed for all of the molecules, basically two paths exist for determining a lower-dimensional coordinate-based representation. In the upper path in Scheme 1 coordinates are determined using either multi-dimensional scaling (MDS) [11] or non-linear mapping



Scheme 1

(NLM) [13] procedures, both of which require minimization of some sort of error function. In the past, both procedures were somewhat limited and could only deal effectively with datasets of less than ~2000 molecules. In addition, they encountered difficulty in treating new sets of compounds that were not included in the original set without redoing the calculations for the entire augmented set. These limitations have been removed by the work of Agrafiotis and his colleagues [100, 101] who developed a clever neural-net approach that learned the non-linear mapping based upon the use of training sets of relatively small sample size (~1,000 compounds). Once the mapping function is learned new compounds can be mapped with relative ease.

The lower path is somewhat more complicated. The first step in the path involves either PCA [12] or principal-coordinate analysis (PCO) [12]. This step can be followed by optimization of a function that minimizes the error between the proximity measure computed in the reduced-dimension and full coordinate systems if desired. Xie et al. [102], recently published an interesting paper along these lines. Kruscal stress [103] is a widely used function in this regard, namely,

$$K_{\text{stress}} = \sqrt{\frac{\sum_i \sum_{j>i} \left(\hat{d}_{i,j} - d_{i,j}\right)^2}{\sum_i \sum_{j>i} \hat{d}_{i,j}^{\,2}}}, \tag{78}$$

where $\hat{d}_{i,j}$ is the distance computed in the reduced-dimension space and $d_{i,j}$ is the distance computed in the full space.

PCA is designed to deal directly with correlation matrices, but not directly with similarity or distance matrices. However, as pointed out by Kruscal [103], the similarity matrix (or other proximity matrix) can be treated as a normal data matrix upon which principal component analysis is performed, that is

$$\mathbf{S} \underset{\substack{\text{mean}\\ \text{center}\\ \text{columns}}}{\Rightarrow\Rightarrow} \bar{\mathbf{S}} \underset{\substack{\text{form}\\ \text{matrix}\\ \text{product}}}{\Rightarrow\Rightarrow} \bar{\mathbf{S}}^{\mathrm{T}}\bar{\mathbf{S}} \underset{\substack{\text{diagonalize}\\ \text{matrix}}}{\Rightarrow\Rightarrow} \mathbf{V}^{\mathrm{T}}\left(\bar{\mathbf{S}}^{\mathrm{T}}\bar{\mathbf{S}}\right)\mathbf{V} = \Lambda, \tag{79}$$

where the columns of the eigenvector matrix $\mathbf{V}$ are the principal components and the elements of the diagonal matrix Λ are the corresponding eigenvalues. The coordinates in the transformed PC coordinate system, usually called the "scores," are given by the matrix $\mathbf{T}$, where

$$\mathbf{T} = \bar{\mathbf{S}}\,\mathbf{V}. \tag{80}$$

Principal coordinate analysis [12] works in an analogous fashion except that the similarity matrix is used directly without the additional multiplications given in Eq. (79). Gower has described

the relationship between PCA and PCO [104]. Because both approaches utilize matrix diagonalization procedures, the size systems that they can practically treat are limited to ~2,000 molecules. This computational obstacle can be overcome for PCA using one of the neural net methods for determining principal components [105]. Benigni [106] described an analogous method based upon a matrix of Euclidean distances computed from high-dimensional vectors representing a set of molecules. Analogous dissimilarity-based methods have also been developed.

An important question is whether the proximity measures are compatible with those of these references addresses the important issue of whether the proximity measure is compatible with embedding in a Euclidean space. For example, satisfying the distance axioms do not guarantee that any distance matrix associated with a given set of molecules will be compatible, as the distance axioms are still satisfied in non-Euclidean spaces. Gower has written extensively on this important issue, and his work should be consulted for details [107–109]. Benigni [110] and Carbó [73] have also contributed interesting approaches in this area.

More recent work by a number of authors has further addressed the issue of embedding of high-dimensional data into lower dimensional spaces. These methods include the similarity-based abstract vector-space approach of Raghavendra and Maggiora [95], isometric mapping [111], local linear embedding [112], exploratory projection-pursuit [113], stochastic proximity embedding [114, 115], and eigenvalue-based methods [116].

3.3. Activity Cliffs and the Topography of Activity Landscapes

Chemical spaces and the activities of molecules in these spaces induce *activity landscapes*. In three dimensions, activity landscapes can be visualized as surfaces with features that are analogous to the Earth's topographical features – mountains, canyons, hills, valleys, cliffs, ridges, spires, plains, etc. Neglecting the Earth's curvature, which for small surface regions can be considered essentially flat, topographies can be represented by two rectilinear position coordinates and one rectilinear altitude coordinate. In activity landscapes, the two position coordinates give the location of a molecule within a two-dimensional chemical space and the altitude coordinate corresponds to the molecule's activity value with respect to a given biological or pharmacological assay. Thus, a number of different activity landscapes exist for a set of molecules in a given chemical space, one for each assay. Because many assays are of relatively low resolution, their activity landscapes will typically be of low resolution. Moreover, since chemical spaces are not invariant to representation, the nature of their associated activity landscapes will also be affected by the representation used to define the chemical space (*Cf.* Fig. 8). Lastly, chemical spaces are typically greater than two dimensions so the geographical analogy associated with the Earth, while familiar, is definitely a significant

simplification of the situation encountered in chemical spaces. Nevertheless, the information provided by simple 3-D models of activity landscapes still provides many useful insights that facilitate our understanding of the actual multi-dimensional case.

Until recently, it was generally assumed that small changes in activity are typically associated with small changes in molecular similarity [4]. In such cases, activity landscapes tend to resemble the rolling hills of Kansas. However, a growing number of studies have shown that this is not generally the case. In fact, activity landscapes appear to contain regions with "cliffs" that resemble the rugged landscape of Bryce Canyon more than that of Kansas [117]. Such "activity cliffs" arise when small changes in molecular similarity result in correspondingly large changes in activity. A recent editorial by Maggiora [118] suggests that activity cliffs play a significant role in the determination of quantitative structure–activity relations (QSAR). This editorial was followed up by several papers that provided additional discussion of this topic [119–121]. *Importantly, activity cliffs pinpoint regions of activity landscapes that contain maximum information on SARs.* This is so because small changes in molecular similarity make it easier to identify what features may be responsible for the dramatic shifts observed in activity. In contrast, it is difficult to ascertain just what features may be responsible for large activity shifts observed between two molecules that are very dissimilar. Figure 11, taken from the recent review by Bajorath et al. [122], provides a 3-D example based on data obtained from a set of Cyclooxygenase-2 inhibitors that illustrates the topography of a typical activity landscape and indicates the relationship between an activity cliff and its associated SAR. Note that the coordinates in the two-dimensional chemical space depicted in the figure are obtained by projection from a higher-dimensional chemical space.

3.3.1. Development of Structure–Activity Similarity Maps

Although the 3-D activity landscape portrayed in Fig. 11 has great intuitive appeal, it does not present an accurate picture of the true activity landscape, which is of much higher dimension. What is needed is a relatively simple representation of the data that can be visualized and analyzed and that captures a significant portion of the information contained in the activity landscape. Early work in this area by Shanmugasundaram and Maggiora [123] introduced the concept of a structure–activity–similarity map or *SAS map*. Figure 12 provides an example of such a map. The ordinate represents the "activity similarity," which is defined by

$$S_{\mathrm{Act}}(\mathrm{A},\mathrm{B}) = 1 - \frac{|\mathrm{Act}(\mathrm{A}) - \mathrm{Act}(\mathrm{B})|}{\mathrm{Act}_{\max} - \mathrm{Act}_{\min}}, \qquad (81)$$

where Act(X) is the activity of compound A or B, typically given in terms of their pI_{50} values, and $\mathrm{Act}_{\max} - \mathrm{Act}_{\min}$ is the difference between the maximum and minimum activity values of the

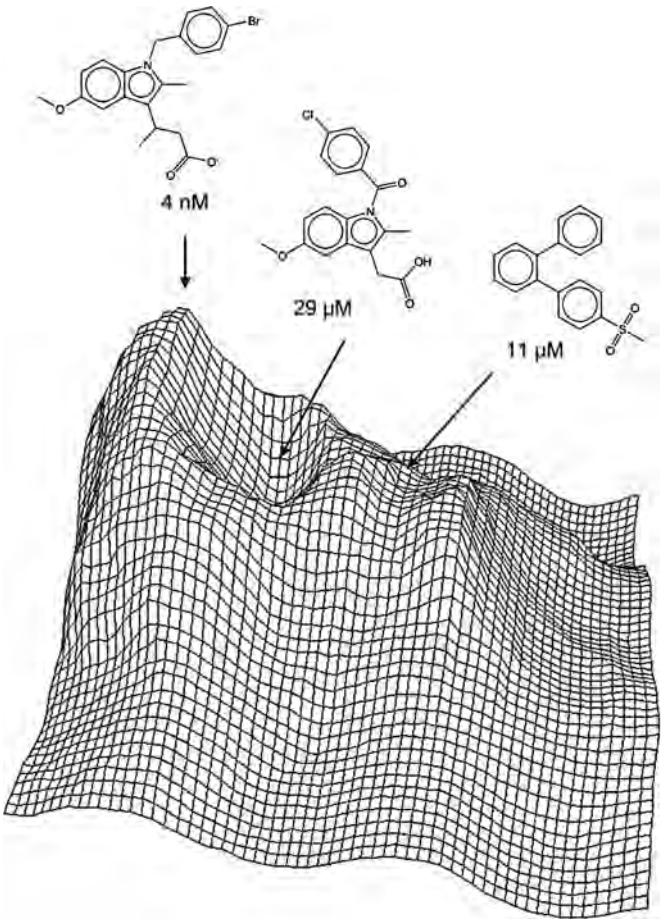


Fig. 11. Example of an activity landscape generated using data obtained from a set of cyclooxygenase-2 inhibitors (*Original figure provided courtesy of Prof. Jurgen Bajorath, see* ref. [122]). Note that the reference chemical space depicted here is a projection onto two dimensions of the actual higher-dimensional chemical space.

compounds in the dataset. This represents a "normalized" activity difference, although non-normalized activity differences can also be used in SAS maps. The abscissa represents the familiar "structure similarity"; any of the similarity measures discussed in this chapter can be used here.

Each of the 1,275 datapoints shown in Panel (a) of the figure represents a pairwise comparison with respect to the activity and structure similarity of each of the molecules in a small, prototypical dataset containing 51 molecules (*N.B.* that the number of distinct pairs obtained from N molecules is $N(N-1)/2$). The data points are color-coded by the activity of the most active compound of a given pair. Red circles denote pairs where at least one compound is active; yellow circles indicate pairs where at least one compound is moderately active; blue circles indicate pairs

where both compounds are inactive (or have low activity). Note that the scale for "Structure Similarity" does not run from zero to unity. This is because of the limited size and diversity of the dataset, since the minimum similarity between any pair of compounds is 0.58.

The SAS map is divided into four quadrants as shown in Panel (b) of the figure. Points located in the upper right quadrant (I) of the diagram correspond to pairs of compounds with both high activity and high structure similarity. The topography in this quadrant is relatively smooth, gently rolling hills. Compounds in this region behave in a "traditional" SAR fashion and reliable QSARs can generally be obtained from such compounds. Points in the lower right quadrant (II) correspond to pairs of compounds with high structure but low activity similarity. This region is rich in activity cliffs, and hence, provides the maximum amount of SAR information. However, compounds in this region tend to be refractory to the determination of reliable QSARs because functions describing QSARs of such compounds must be highly flexible to account for the high variability of their activity landscapes. Thus, a considerable amount of SAR data is required to ensure that the functions, which are highly non-linear, are adequately approximated. The upper left quadrant (III) corresponds to pairs of compounds with high activity similarity but low structure similarity. Compounds in this region exhibit relatively low *local* SAR information, although they do provide SAR information on *distributed* classes of active compounds (*Cf.* "scaffold hopping" [124, 125]). Because of the low similarities of compounds within this quadrant (*N.B.* that compounds in this set are typically dispersed in chemical space) it is generally not possible to develop reliable QSARs. However, if enough SAR data is available for compounds in both "active classes," it may be possible to develop QSARs for each class separately. The separate QSARs can then be merged into a single "distributed" QSAR. Care must be exercised in interpreting the points in quadrants (II) and (III), since high activity similarity obtains when pairs compounds have similar activities that can be high, moderate, low, or inactive. This can be indicated in SAS maps by distinguishing pairs of compounds using a color-coding scheme such as that described above for Fig. 12. Lastly, the lower left quadrant (IV) corresponds to pairs of compounds with both low activity and low structure similarity. This region contains very little if any SAR information and, thus, compounds in this region do not submit to QSAR or provide useful information on the nature of the activity landscape.

Figure 13 depicts the same SAS map shown in Fig. 12. Two points, one located in quadrant (II) and one in quadrant (III) are explicitly indicated by the green arrows that point from pairs of molecules located in Boxes (A) and (B), respectively. The point in quadrant (II) corresponds to an activity cliff since the activities of

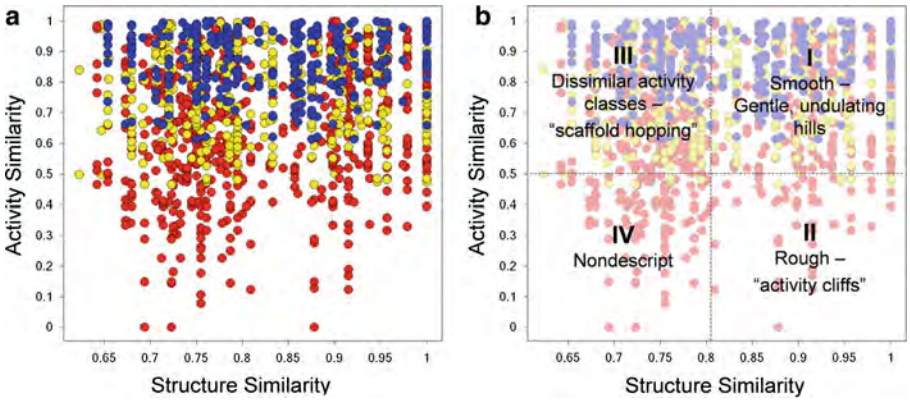


Fig. 12. Example of a structure–activity similarity (SAS) map (*Original figure provided courtesy of Dr. José Medina-Franco*). (**a**) Depiction of a SAS map for a prototypical dataset. Each data point indicates a pairwise comparison from a dataset of 51 compounds. Data points are color coded by the activity of the most active compound of a given pair. Red circles denote pairs where at least one compound is active; yellow circles indicate pairs where at least one compound is moderately active; blue circles indicates pairs where both compounds are inactive (or have low activity). Note that the scale for 'Structure Similarity' does not run from zero to unity. This is because of the limited size and diversity of the dataset. (**b**) Depiction of the four approximate quadrants of the SAS map (*see* text for further discussion).

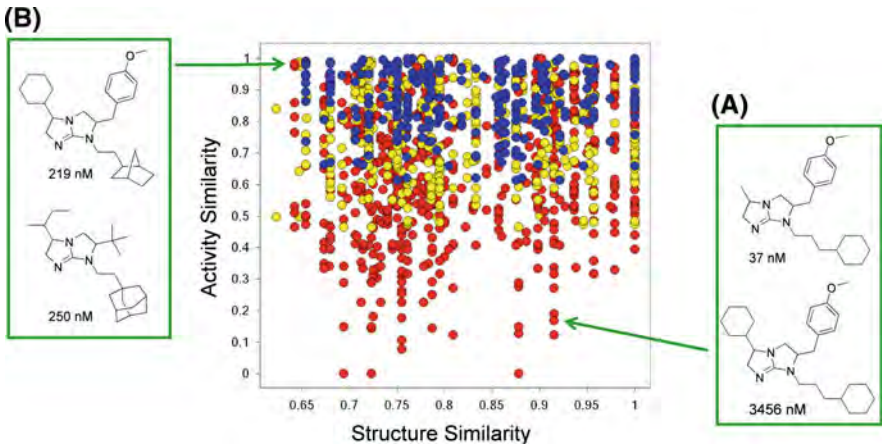


Fig. 13. SAS map given in Fig. 12 showing the structures of a pair of compounds involved in an activity cliff and a pair of compounds involved in scaffold hopping (*Original figure provided courtesy of Dr. José Medina-Franco*). The pair of compounds in Box (A) lie in quadrant (II), and the pair of compounds in Box (B) lie in quadrant (III) of the SAS map (*see* also Fig. 12b).

the two compounds in Box (A) differ by a factor of nearly 100, while their structure similarity is greater than 0.9. The two compounds associated with the designated point in quadrant (III) have moderate and nearly equal activities. However, since the dataset is of rather limited diversity – the smallest similarity between any two compounds in the set is 0.58 (*vide supra*) – it is not expected that the compounds associated with this datapoint will exhibit significant scaffold hopping, which is confirmed by the

structures in Box (B). It is expected that noteworthy scaffold hopping will require similarity values most likely of 0.40 or less. Thus, one might only expect this to occur in campaigns where relatively diverse sets of compounds are screened.

3.3.2. An Information-Theoretic Analysis of SAS Maps

Information theory provides a suitable framework for analyzing the information content of pairs of compounds located in the different quadrants of SAS maps, such as that depicted in Fig. 12b. Information (in "bits"), sometimes called "surprisal," is related to Shannon entropy [126] and is given by

$$\mathcal{I}(\mathrm{A}) = \log_2 \frac{1}{P(\mathrm{A})} = -\log_2 P(\mathrm{A}), \tag{82}$$

where $P(\mathrm{A})$is the probability of observing a specific activity (e.g., high, moderate, low, etc.) activity for molecule A. The condensed notation employed here is used for simplicity. Technically, $P(\mathrm{A})$ should *conditioned* on the activity of molecule B and the similarity, $S(\mathrm{A,B})$, of the two molecules. For example, what is the probability that molecule A is active (inactive) given that molecule B is active (inactive) and the similarity between them is high (low). Equation (82) makes sense from the following point of view, namely, the more likely an event is to be observed, the less information will be obtained in observing it, that is there is less "surprise" in observing it. For example, if an urn is filled with 90 red balls and 10 green balls, there is a 90% chance of drawing a red one and a 10% chance of drawing a green one. Thus, drawing a red ball is less surprising than drawing a green one and hence carries less information.

Although the number of exceptions to the rule that "similar compounds tend to possess similar activities" [4] is growing rapidly, it is not unreasonable to assume, at least as a working hypothesis, that this rule holds approximately. A compound located in the neighborhood of, say, an active compound is likely to also be active. If this is not the case, and the compound is found to be inactive, the result is surprising. Hence, such an observation carries higher information than if the compound was active as expected. Using this logic and the basics of information theory it is possible to *qualitatively* assess the degree of information possessed by pairs of compounds located within the different quadrants of a SAS map.

For example, consider two highly similar compounds chosen at random from a given chemical space. If one of the compounds is known to be active, then it is reasonable to expect that the other would have a high probability of also being active. This situation is exemplified by compounds residing in quadrant (I). Thus, although compounds in this quadrant are appropriate for the development of reliable QSARs, they provide little information, in the information-theoretic sense. If, on the other hand, one of the compounds is active but the other is inactive, a condition

exemplified by the activity cliffs associated with compounds in quadrant (II), this corresponds to a high information local case. However, due to the rapidly fluxuating nature of the activity landscape for compounds located in quadrant (II) it is difficult to construct meaningful QSARs (*see* Subsection 3.3.1 for further discussion). If two dissimilar compounds are chosen at random, and one is known to be active, it follows from the above discussion that the other is likely to be inactive. However, if it is shown to be active, a situation that obtains for compounds located in quadrant (III), this corresponds to a high-information case since a new class of compounds has been identified. Lastly, pairs of compounds that have low activity and structure similarity and are found in quadrant (IV) have little relationship to each other and, hence, have low information.

The above discussion can also be couched in terms of inactive compounds. This is obtained by interchanging the words "active" and "inactive." In such a case, inactive compounds in quadrants (I) and (III) have low information and are also not appropriate for QSAR studies. Table 1 provides a summary of the features of SAS maps.

3.3.3. Alternative Representations of Activity Landscapes

Many variants of SAS maps are possible. One that has received some attention recently is the multi-fusion similarity (MFS) maps developed by Medina-Franco et al. [127]. This work is a two-dimensional extension of the work carried out by Willet's group on similarity-based data fusion methods [25–28].

Several other approaches aimed at describing activity landscapes have been published [128, 129]. These approaches are

Table 1
Structure–activity similarity (SAS) maps

Quad	Similarity		Landscape	Cpd activity[a]	QSAR[b]	Information content
	Activity	Structure				
I	High	High	Gentle hills	High/high Low/low	+ –	Low
II	Low	High	Activity cliffs	High/low	–	High
III	High	Low	Multiple, separated active regions	High/high Low/low	–/+ –	High to moderate
IV	Low	Low	Nondescript	High/low	–	Very low

[a]If a pair of compounds has high activity similarity, both of the compounds could have high or low activities. This has been explicitly designated (*viz.*, High/High or Low/Low) for compounds in quadrants (I) and (III)

[b]The "+" sign indicates that it is possible to construct a reasonable QSAR; the "–" sign indicates that a QSAR is either not possible or problematic at best; the "–/+" indicates that it is a QSAR is not possible unless a significant amount of data is available for both of the dissimilar activity classes

based on two indices – SALI [128] and SARI [129]. The former, which is defined as

$$\mathrm{SALI}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{Act}(\mathrm{A}) - \mathrm{Act}(\mathrm{B})|}{1 - S(\mathrm{A},\mathrm{B})}, \tag{83}$$

is designed to identify the presence of activity cliffs between pairs of compounds. Thus, it provides a *local* characterization of the neighborhoods surrounding activity cliffs. A more global view is obtained by stitching together compounds using a directed graph-theoretical formalism. In this formalism, the nodes represent compounds. A *directed* edge is drawn between two compounds and points towards the higher activity one depending upon whether the SALI index is above a given threshold value. This representation provides a graphical means for portraying neighborhood as well as global relationships among compounds associated with activity cliffs – as the SALI threshold value is lowered a more complete picture of the inter-relationships among activity cliff regions emerges.

In contrast, the SARI index takes a more global approach to activity landscapes. It is based on the average of the "continuity" score, $\langle \mathrm{Score}_{\mathrm{cont}} \rangle$, and the "discontinuity" score $\langle \mathrm{Score}_{\mathrm{discont}} \rangle$, The former is computed by taking the potency-weighted sum of all pairwise dissimilarities and, thus, is a measure of the potency and diversity of the set of compounds under consideration. In contrast, the latter is computed by taking similarity-weighted average potency among pairs of compounds that exceed a given similarity threshold value. Large values of the discontinuity score indicate the presence of activity cliffs. After normalizing both scores to the unit interval, they are combined as shown in Eq. (84) to yield the SARI index value,

$$\mathrm{SARI} = \tfrac{1}{2}\left[\langle \mathrm{Score}_{\mathrm{cont}} \rangle_{\mathrm{norm}} + \left(1 - \langle \mathrm{Score}_{\mathrm{discont}} \rangle_{\mathrm{norm}}\right)\right]. \tag{84}$$

High SARI values correspond to predominantly continuous activity landscapes, while low values correspond to predominantly discontinuous landscapes. Intermediate values correspond to mixed landscapes. Hence, the SARI index provides a more global measure of the activity landscape than does the SALI index. However, it provides a much less detailed picture of the local environments in an activity landscape, although a local discontinuity index has also been defined for the former [129]. Subsequent work in Bajorath's laboratory [130] has extended their SARI-based approach using network-like similarity graphs (NSG) along with local information to provide a more detailed picture of activity landscapes.

Lastly, since similarity values are influenced by the representation used to encode the molecular information, it is desirable to develop a method that is less sensitive to representation. To

address this difficult and persistent problem, Medina-Franco *et al.* [131] developed the concept of *consensus activity cliffs*. Consensus activity cliffs, which are obtained by analyzing multiple activity landscapes generated by different similarity methods, are characterized by the Degree of Consensus (DoC) between any pair of similarity methods. Using this approach, the authors were able to identify activity cliffs of improved reliability that persisted with respect to a number of similarity methods.

3.4. A General Similarity-Based Approach for Representing Chemical Spaces

Vector-based representations of chemical spaces are quite common in cheminformatics. In the usual molecular fragment-based approach the coefficients of the vector components are generally binary- or positive integer-valued (*see* Subsections 2.2.1 and 2.2.3). In continuous vector representations (*see* Subsection 2.3), on the other hand, the vector coefficients are typically associated with the values of atomic and molecular properties. All of these representations can be used to describe chemical spaces, either directly or in terms of their related molecular similarities, proximities, or distances.

In the following, a general similarity-based approach for constructing continuous vector representations of molecules is presented that is based on what might be called "molecular basis vectors" instead of molecular fragments or properties (*Cf.* Subsection 2.3). The method is reminiscent of those in molecular quantum mechanics [132] that employ atomic orbitals or group functions as a basis for describing whole molecules. A detailed account with a number of examples can be found in a recent work by Raghavendra and Maggiora [95]. As will be seen in the sequel, the key to the method is an *ansatz* that equates the inner product between a pair of molecular vectors to their corresponding similarities. The method's power resides in the fact that any reasonable similarity can be used and that the detailed nature of the molecular vector need not be known since it is not required for the computation of similarity. This situation is reminiscent of that in many kernel learning methods [133–135] where elements of the Gramian matrix, which are analogous to the similarities used here, can be determined indirectly without knowledge of the explicit form of the vectors in the underlying feature space.

Choose a set of p molecules as a molecular basis for representing the chemical space of interest

$$\mathrm{B} = \{b_1, b_2, ..., b_p\}, \tag{85}$$

which can be written as the "row vector" (i.e. a $1 \times n$-dimensional matrix)

$$\mathrm{B} = (\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_p), \tag{86}$$

where the $\mathbf{b}_i$, $i = 1, 2, ..., p$ correspond to *molecular basis vectors*. Here, the word "vector" refers to an abstract object with direction

and magnitude located at the origin of an appropriate coordinate system and satisfying the multiplicative and additive properties of a linear vector space [36]. Such objects, which are depicted in a lower-case, bold-face Arial font, e.g. "$\mathbf{b}_i$, $\mathbf{m}_k$, ...", are *basis-set independent*. Component vectors, which are *basis-set dependent*, are depicted in a lower-case, bold-face Times New Roman font, e.g. "$\mathbf{v}_i$, $\mathbf{m}_k$, ...". Their components are depicted in lower-case italic type and are grouped together in $n \times 1$ column matrices (*see* Eq. (91)).

Consider the similarities of all of the elements of the molecular basis set with respect to each other. This formally generates the matrix of inner products [136]

$$\mathbf{S} = \langle \mathbf{B}, \mathbf{B} \rangle = \begin{pmatrix} \langle \mathbf{b}_1, \mathbf{b}_1 \rangle & \langle \mathbf{b}_1, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_1, \mathbf{b}_p \rangle \\ \langle \mathbf{b}_2, \mathbf{b}_1 \rangle & \langle \mathbf{b}_2, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_2, \mathbf{b}_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{b}_p, \mathbf{b}_1 \rangle & \langle \mathbf{b}_p, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_p, \mathbf{b}_p \rangle \end{pmatrix}, \tag{87}$$

$$\mathbf{S} = \begin{pmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,p} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p,1} & S_{p,1} & \cdots & S_{p,p} \end{pmatrix}. \tag{88}$$

Since the off-diagonal elements are, in general, non-zero and since the diagonal elements are of unit value the molecular basis vectors constitute a set of non-orthogonal unit vectors. A crucial feature of this approach is that the exact nature of the molecular basis vectors need not be known. Due to the *ansatz*, only their inner products are required, and these are taken to be equivalent to the similarities among pairs of the corresponding molecules (*vide supra*). Since $0 < S_{i,j} \leq 1$, the elements of **S** are analogous to the basis-set overlaps integrals familiar in quantum chemistry [69]. Moreover, **S** is positive definite if all of the elements of the molecular basis are linearly independent; when they are not, **S** becomes positive semi-definite. Hence, the definiteness of the **S** matrix provides a measure of the linear independence of the molecular basis set.

Similarities between molecular basis elements can be evaluated in a number of different ways. For example, suppose the $\mathbf{b}_i \Leftrightarrow \mathrm{G}_i$, that is the i-th element of the molecular basis is a *labeled chemical graph* of a molecule. Then $S_{i,j} = \langle \mathbf{b}_i, \mathbf{b}_j \rangle \equiv S_{\mathrm{Tan}}(G_i, G_j)$, where the Tanimoto similarity, S_{Tan}, is evaluated as in Eq. (15). The set of labeled graphs $G = \{G_1, G_2, ..., G_p\}$ can be referred to as a "chemical graph basis". Similarities can also be computed using a bit-vector representation or from their 3-D structures or molecular fields as described in the preceding sections. In some

applications **S** is equivalent to what is typically called the *metric matrix*; in statistics **S** is equivalent to the *correlation matrix* [136].

Consider a given molecule m_i within a set of n molecules

$$\mathrm{M} = \{m_1, m_2, ..., m_n\}. \tag{89}$$

A molecule $m_i \in \mathrm{M}$ generally does not correspond to any of the basis molecules in B, although such a correspondence is not specifically precluded on mathematical grounds because of the non-orthogonality of the molecular basis. In matrix notation, a given molecule $m_i \in \mathrm{M}$, can be represented as a abstract vector m_i in the molecular basis,

$$\mathbf{m}_i = \mathbf{B}\,\mathbf{m}_i, \tag{90}$$

where the column vector of coefficients is given by

$$\mathbf{m}_i = \begin{pmatrix} m_i(b_1) \\ m_i(b_2) \\ \vdots \\ m_i(b_p) \end{pmatrix}. \tag{91}$$

To compute the various cosine-like similarity indices it is necessary to evaluate the inner product $\langle \mathbf{m}_i, \mathbf{m}_j \rangle$ and vector norm $\| \mathbf{m}_i \| = \sqrt{\langle \mathbf{m}_i, \mathbf{m}_i \rangle}$ (*see* also Eqs. (45) and (46)):

$$\begin{aligned} \langle \mathbf{m}_i, \mathbf{m}_j \rangle &= \langle \mathbf{B}\,\mathbf{m}_i,\ \mathbf{B}\,\mathbf{m}_j \rangle \\ &= \mathbf{m}_\mathbf{i}^{\mathrm{T}} \langle \mathbf{B},\ \mathbf{B} \rangle\, \mathbf{m}_j \end{aligned}. \tag{92}$$

In expanded form, the inner product is given by

$$\begin{aligned} \langle \mathbf{m}_i, \mathbf{m}_j \rangle = &\left(m_i(b_1), m_i(b_2), ..., m_i(b_p) \right) \\ &\times \begin{pmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,p} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p,1} & S_{p,1} & \cdots & S_{p,p} \end{pmatrix} \begin{pmatrix} m_i(b_1) \\ m_i(b_2) \\ \vdots \\ m_i(b_p) \end{pmatrix}. \end{aligned} \tag{93}$$

In summation form, Eq. (93) becomes

$$\langle \mathbf{m}_i, \mathbf{m}_j \rangle = \sum_{k=1}^{p} \sum_{\ell=1}^{p} m_i(b_k) \cdot m_j(b_\ell) \cdot S_{k,\ell}, \tag{94}$$

where $S_{i,i} = 1$ for $i = 1, 2, ..., p$. Comparing Eq. (94) with Eq. (45) shows that the elements of the S-matrix modulate the product of the vector components and the cross-terms, "$m_i(b_k) \cdot m_j(b_\ell)$", remain. When the basis is orthonormal $\mathbf{S} = \mathbf{I}$ and Eq. (94) reduces to Eq. (45). Similarly, the vector norm for the i-th molecule is given by

$$\|\mathbf{m}_i\| = \sum_{k=1}^{p} \sum_{\ell=1}^{p} m_i(b_k) \cdot m_i(b_\ell) \cdot S_{k,\ell}, \tag{95}$$

which reduces to Eq. (46) when $\mathbf{S} = \mathbf{I}$. These relationships clearly show the important role played by the metric matrix **S**. Since the various cosine-like similarity indices all depend on the quantities given in Eqs. (94) and (95), it follows that these indices also depend upon **S**, but this dependence is routinely neglected in most calculations. Euclidean (*see* Eq. (42)) and other distances are likewise affected by the metric matrix:

$$\begin{aligned} d_{\text{Euc}}(\mathbf{m}_i, \mathbf{m}_j) &= \left\| \mathbf{m}_i - \mathbf{m}_j \right\| \\ &= \sqrt{\left\langle (\mathbf{m}_i - \mathbf{m}_j), (\mathbf{m}_i - \mathbf{m}_j) \right\rangle} \\ &= \sqrt{\sum_{k=1}^{p} \sum_{\ell=1}^{p} \left(m_i(b_k) - m_j(b_\ell) \right) \cdot \left(m_i(b_k) - m_j(b_\ell) \right) \cdot S_{k,\ell}} \end{aligned} . \tag{96}$$

As was true in the two cases above, Eq. (96) reduces to Eq. (42) in an orthonormal basis.

There are numerous ways in which to orthonormalize a basis [137]. Here, we choose to employ the *symmetric orthonormalization* procedure described by Löwdin [136]

$$\bar{\mathbf{B}} = \mathbf{B}\,\mathbf{S}^{-\frac{1}{2}}, \tag{97}$$

where

$$\bar{\mathbf{B}} = (\bar{\mathbf{b}}_1, \bar{\mathbf{b}}_2, \ldots, \bar{\mathbf{b}}_p). \tag{98}$$

This has the benefit over other orthogonality procedures that the new basis is as close as possible, in a least square sense, to the original basis [138]. Computing the inner product, $\left\langle \bar{\mathbf{B}}, \bar{\mathbf{B}} \right\rangle = \left\langle \mathbf{B}\,\mathbf{S}^{-\frac{1}{2}}, \mathbf{B}\,\mathbf{S}^{-\frac{1}{2}} \right\rangle = \mathbf{S}^{-\frac{1}{2}} \langle \mathbf{B}, \mathbf{B} \rangle\, \mathbf{S}^{-\frac{1}{2}} = \mathbf{S}^{-\frac{1}{2}}\,\mathbf{S}\,\mathbf{S}^{-\frac{1}{2}} = \mathbf{I}$, shows that the basis is indeed orthonormal.

Right multiplying the terms in Eq. (97) by $\mathbf{S}^{\frac{1}{2}}$ gives $\mathbf{B} = \bar{\mathbf{B}}\,\mathbf{S}^{\frac{1}{2}}$, which upon substitution into Eq. (90) yields

$$\begin{aligned} \mathbf{m_i} &= \mathbf{B}\,\mathbf{m}_i \\ &= (\bar{\mathbf{B}}\,\mathbf{S}^{\frac{1}{2}})\mathbf{m}_i = \bar{\mathbf{B}}\,(\mathbf{S}^{\frac{1}{2}}\,\mathbf{m}_i) \\ &= \bar{\mathbf{B}}\,\bar{\mathbf{m}}_i \end{aligned} , \tag{99}$$

where the "expansion coefficients" (i.e. components) in the new, orthonormal basis are given by

$$\bar{\mathbf{m}}_i = \mathbf{S}^{\frac{1}{2}}\,\mathbf{m}_i. \tag{100}$$

As was the case for the basis above, Eq. (100) can be rearranged to give the expansion coefficients in the original, non-orthogonal basis,

$$\mathbf{m}_i = \mathbf{S}^{-\frac{1}{2}}\,\mathbf{m}_i. \tag{101}$$

Thus, this equation provides the means for determining the components of $\mathbf{m}_i$ in the original basis given the components in the orthonormal basis, which are easily determined. This can be accomplished by first taking the inner product of the k-th orthonormal basis element with $\mathbf{m}_i$ (*see*, e.g., Eq. (94))

$$\begin{aligned}\left\langle \bar{\mathbf{b}}_k, \mathbf{m}_i \right\rangle &= \left\langle \bar{\mathbf{b}}_k, \bar{\mathbf{B}}\, \bar{\mathbf{m}}_i \right\rangle \\ &= \sum_{\ell=1}^{p} \left\langle \bar{\mathbf{b}}_k, \bar{\mathbf{b}}_\ell \right\rangle \bar{m}_i(\bar{b}_\ell).\end{aligned} \tag{102}$$

Since $\left\langle \bar{\mathbf{b}}_k, \bar{\mathbf{b}}_\ell \right\rangle = \delta_{k,\ell}$, where the Kronecker delta, $\delta_{k,k} = 1$ and $\delta_{k,\ell} = 0$ for $k \neq \ell$ the k-th component of $\mathbf{m}_i$ is given by

$$\bar{m}_i(\bar{b}_k) = \left\langle \bar{\mathbf{b}}_k, \mathbf{m}_i \right\rangle \quad \text{for } k = 1, 2, ..., p. \tag{103}$$

Because $\mathbf{m}_i$ is normalized with respect to the Euclidean norm

$$\sum_{k=1}^{p} \bar{m}_i(\bar{b}_k)^2 = 1 \tag{104}$$

the *square* of each component value, $\bar{m}_i(\bar{b}_k)$, gives the fraction of the molecule represented by its corresponding orthonormal basis element $\bar{\mathbf{b}}_k$.

To evaluate the inner product $\left\langle \bar{\mathbf{b}}_k, \mathbf{m}_i \right\rangle$, $\bar{\mathbf{b}}_k$ must be expanded in terms of the original non-orthogonal basis (*see* Eq. (97)), that is

$$\bar{\mathbf{b}}_k = \sum_{\ell=1}^{p} \mathbf{b}_\ell\, S_{k,\ell}^{-\frac{1}{2}}. \tag{105}$$

Substituting Eq. (105) into Eq. (103) yields

$$\bar{m}_i(\bar{b}_k) = \sum_{\ell=1}^{f} S_{k,\ell}^{-\frac{1}{2}} \left\langle \mathbf{b}_\ell, \mathbf{m}_i \right\rangle. \tag{106}$$

The inner-product terms $\left\langle \mathbf{b}_\ell, \mathbf{m}_i \right\rangle$ can now be evaluated in exactly the same manner as was described earlier using the chosen similarity measure. For example, $\left\langle \mathbf{b}_\ell, \mathbf{m}_i \right\rangle = \mathrm{S_{Tan}}(\mathrm{G}_\ell, \mathrm{G}_m)$, where "$\mathbf{b}_\ell$" is the labeled graph corresponding to ℓ-th basis molecule, "$\mathbf{m}_i$" is the labeled graph corresponding to the i-th molecule, and $S_{\mathrm{Tan}}(G_\ell, G_m)$ is the chemical graph-theoretical Tanimoto similarity coefficient.

This approach can, in many instances, be extended even to cases where the basis is comprised of physico-chemical, topological, or other such parameters. The similarity matrix is replaced in these cases by the correlation matrix computed with respect to the "basis set" of parameters.

Agrafiotis et al. [139], developed a similar approach to generate vectors for input into neural nets. Although these authors did not account for the inherent non-orthogonality of the "basis," in their work, the issue of the orthogonality of the basis may be less critical than it is here, and the mappings they generated seem to

be sufficiently stable. Another related approach comes from Villar and co-workers [140]. In this case, however, the basis consisted of a set of proteins. The interaction of each molecule in the training set to each of the proteins in the "basis proteins" was measured experimentally, and the expansion coefficients were determined using a least squares procedure. Again, non-orthogonality of the basis was not explicitly addressed, although the choice of the basis proteins did involve an assessment of correlations among them.

Randic [141–143] has investigated the role of orthogonalized descriptors in multivariate regressions. In his work he points out that although the predictions obtained with orthogonal or non-orthogonal descriptors are the same, the stability of the regression coefficients is much greater in the former case. Also, adding a new, orthogonal descriptor to set of orthogonal descriptors does not affect the values of the previously determined regression coefficients. This is definitely not the case for non-orthogonal descriptors where addition of a new descriptor can cause all of the coefficients to fluctuate significantly depending on the degree of collinearity of the new descriptor with those in the original set.

4. Summary and Conclusions

This chapter provides an overview of the mathematics that underlies many of the similarity measures used in cheminformatics. Each similarity measure is made up of two key elements: (1) A mathematical representation of the relevant molecular information and (2) some form of similarity measure, index or coefficient that is compatible with the representation. The mathematical forms typically used are sets, graphs, vectors, and functions, and each is discussed at length in this chapter.

As was described in Subsection 2.1, chemical graphs are a subclass of mathematical graphs, and thus many of the features of the latter can be taken over to the former. A number of graph metrics, such as the size of a graph and the distance between two graphs, have been applied to chemical graphs. In addition, similarity measures, such as the Tanimoto similarity index, also have their corresponding graph-theoretical analogs and have been used in a number of cases, albeit on relatively small sets of molecules. Although chemical graphs are the most familiar and intuitive representation of molecular information to chemists, they have been used relatively rarely in MSA. This is due primarily to computational difficulties brought on by the need to evaluate the MCS, an NP-complete computational problem that is required by most graph-based distance and similarity measures.

Subsection 2.2 describes the properties of discrete-valued feature vectors, with components given by finite, ordered sets of

values. The most prevalent class is that of vectors with binary-valued components, which are mathematically equivalent to classical sets. Here features are either in the set (component value of "1") or not in the set (component value of "0"). Because we are essentially dealing with sets, the distance and similarity measures used are typically related to set measures (i.e. cardinalities) and not to the types of inner (scalar) products defined on linear vector spaces. A hypercubic mathematical space associated can be associated with classical sets, where the dimension of the space is equal to the number of elements in the universal set and each vertex of the hypercube corresponds to a subset, including the null and universal sets. Distances in these spaces are appropriately Hamming distances that satisfy an ℓ_1 metric. Although Euclidean distances are sometimes used, they are inappropriate in such hypercubic spaces. Most similarity indices are taken to be symmetric ("A is as similar to B as B is similar to A"), but Tversky defined an infinite family of asymmetric indices related to the Tanimoto similarity index, some of which may be useful for similarity-related tasks such as similarity searching.

Another class of discrete-valued feature vectors useful in MSA is integer- and categorical-valued feature vectors. Here, the vectors are mathematically equivalent to multisets and not directly to classical sets, although multisets can be reformulated as classical sets. The components of the vectors now indicate the number of times a given feature occurs or the ordered set of categorical values corresponding to the given feature or property. Although care must be taken, distance and similarity measures analogous to those used for binary-valued vector components can be used here as well.

In Subsection 2.3, the important class of vectors with continuous-valued components is described. A number of issues arise in this case. Importantly, since the objects of concern here are vectors, the mathematical operations employed are those applied to vectors such as addition, multiplication by a scalar, and formation of inner products. Care, however, must be exercised because in some cases the "vector objects" may not reside in linear vector spaces. While distances between vectors are used in similarity studies, inner products are the most prevalent type of terms found in MSA. Such similarities, usually associated with the names Carbó and Hodgkin, are computed as ratios, where the inner product term in the numerator is normalized by a term in the denominator that is some form of mean (e.g. geometric or arithmetic) of the norms of the two vectors.

The notion of an orthogonal set of "basis vectors" is also of significance here and is particularly important since as discussed in Subsection 2.3.2 it is in many instances ignored. In a non-orthogonal basis the associated similarity matrix defines the metric of the space in which the vectors "live." Thus, "measurements" such as

the distance or the angle between two vectors in the space are dependent on the metric of that space. Further discussion on this point is presented in Subsection 3.1.3 that describes a general approach for dealing with non-orthogonal bases and explores some of the consequences of ignoring non-orthogonality in the description of chemical spaces. While most of the discussion deals with what are called "molecular basis sets", the method can also deal with physico-chemical, topological, or other such descriptors. However, in these cases the correlation matrix replaces the similarity matrix.

Subsection 2.4 addresses the use of field-based functions in MSA. Field-based functions, which can be thought of as infinite-dimensional vectors, are used primarily in 3-D MSA. Here, molecular fields (e.g. steric or electrostatic) or pseudo-fields (e.g. lipophilic) of the molecules being compared are matched, using various similarity measures, the most popular being those of Carbó or Hodgkin. Because 3-D field-based similarities are non-linear functions, multiple solutions corresponding to different alignments are possible. This has raised the issue of how one obtains consistent multi-molecule consistent alignments, a subject that is treated in Subsection 2.4.3. Conformational flexibility adds a new degree of difficulty to studies of 3-D MSA, and this has been dealt with in a number of ways. The most widespread approach is by standard conformational analysis. Since such an analysis leads to many conformations clustering is usually used to group the conformations as a basis for identifying a smaller set of prototypical conformations. Molecular similarity is then carried out by pairwise matching the fields generated by each conformational prototype in one molecule with each conformational prototype in the other molecule being compared. This represents a rather substantial computational problem that has been ameliorated somewhat using Fourier transforms to separate translational from rotational motions in the optimization process. Alternatively, several procedures have been developed that combine conformational analysis with 3-D similarity matching simultaneously in the optimization process. Both approaches are, however, computationally demanding, although the latter is somewhat better in this regard. Since multiple conformers for each molecule may contribute to the overall similarity, Subsection 2.4.5 deals with a possible way of combining this information into a single multi-conformer dependent similarity.

Subsection 2.5 provides a very brief discussion of molecular dissimilarity measures that are basically the complement of their corresponding molecular similarity measures. This section also presents reasons as to why similarity is preferred over dissimilarity, except in studies of diversity, as a measure of molecular resemblance.

The concept of chemical space pervades, either explicitly or implicitly, much of the literature in cheminformatics. As is

discussed in Subsection 3, chemical spaces are induced by various similarity measures. The different similarity measures do not necessarily give rise to topologically equivalent chemical spaces – nearest-neighbor relations are generally not preserved among chemical spaces induced by different similarity measures. The consequences of this are manifold. An especially egregious consequence is that the results of similarity searches based upon different similarity measures tend can differ substantially. And there is no easy solution to this problem.

Chemical spaces fall into two broad categories, coordinate-based and coordinate-free. Coordinate-based chemical spaces, even those of relatively low dimensionality, tend to be of difficult to visualize directly. Coordinate-free chemical spaces cannot be visualized directly since coordinates do not exist. In both cases it is possible to develop reduced-dimension representations that are easier to work with theoretically and also afford possibilities for visualization. Constructing reduced dimension spaces, which is discussed in Subsection 3.2 for the case of chemical spaces, is a difficult problem that pervades many fields, and methods developed in these fields have proved useful in cheminformatics, albeit to varying degrees.

The growing importance of activity landscapes, and more specifically activity cliffs, are the subject of Subsection 3.3. A growing body of data shows quite clearly that the old aphorism "similar molecules have similar activities" is no longer entirely valid due to the presence of activity cliffs. This has important implications for QSAR studies. However, considerable SAR information is contained in activity cliffs that arise when small changes in similarity lead to large changes in activity. Recently, their growing importance is being recognized, and a number of studies addressing many issues associated with them have been published.

MSA has developed substantially over the years, especially as digital computers became faster, more compact, and widely available to scientists. Handling large sets of molecules is generally not a problem. The main problem confronting MSA is the problem of the lack of topological invariance of the chemical spaces induced by the various similarity measures. Unfortunately, this problem may be fundamentally related to the inherent subjectivity of similarity and thus cannot be addressed in any simple manner.

5. Appendix: A New Notation for Classical Sets

Sets are very general mathematical objects that are used in many branches of mathematics. Here the focus is on *finite* sets, that is sets with a finite set of elements. A key concept in set theory is that of the *universal set*, U, sometimes called the *universe of*

discourse, which is an unordered collection of n elements $x_1, x_2, ..., x_k, ..., x_n$ and is given by

$$\mathrm{U} = \{x_1, x_2, ..., x_k, ..., x_n\}. \tag{107}$$

All sets in this "universe," including U and the null or empty set Ø, are subsets of U. A subset A is typically written as, for example,

$$\mathrm{A} = \{x_1, x_3, x_9, x_{12}, x_{17}, x_{18}, ...\}, \tag{108}$$

but his notation can become awkward and cumbersome for large, complex sets. A more general and powerful notation, which utilizes the concept of an *indicator* or *characteristic function*, $A(x_k)$, is illustrated in Eq. (109),

$$\mathrm{A} = \{A(x_1), A(x_2),..., A(x_k),..., A(x_n)\}, \tag{109}$$

where $A(x_k)$ characterizes the membership of each element in the set is given by

$$A(x_k) = \begin{cases} 1 & \text{if } x_k \in A \\ 0 & \text{if } x_k \notin A \end{cases}. \tag{110}$$

Thus, in the universal set $A(x_k) = 1 \quad \text{for } k = 1, 2, ..., n$, that is all elements of the universal set have a membership-function value of unity.

Note that this representation differs from that usually used (*see* Eq. (108)) where only those elements actually in the set, that is those elements for which $A(x_k) = 1$, are included explicitly. All possible sets A, including the empty and universal sets Ø and U, are subsets of U, i.e. $A \subseteq U$. While this notation may be unfamiliar, it is completely equivalent to that used for binary vectors or "bit vectors." Fuzzy sets, although not treated in this chapter, can also be represented in this notation with the modification that elements of the set are no longer confined to the binary values {0,1}; fuzzy sets can take on all values between and including zero and unity [51]. A number of useful operations between two sets, A and B, are given in the notation introduced above:

$$A \cap B = \min_k [A(x_k), B(x_k)] \quad \text{Set Intersection} \tag{111}$$

$$A \cup B = \max_k [A(x_k), B(x_k)] \quad \text{Set Union} \tag{112}$$

$$\begin{aligned} A^{\mathrm{c}} &= \{1 - A(x_1), 1 - A(x_2), ..., 1 - A(x_n)\} \\ &= \{A^{\mathrm{c}}(x_1), A^{\mathrm{c}}(x_2), ..., A^{\mathrm{c}}(x_n)\} \end{aligned} \quad \text{Set Complementation} \tag{113}$$

$$\begin{aligned} A - B = A \cap B^{\mathrm{c}} &= \min_k [A(x_k), 1 - B(x_k)] \\ &= \min_k [A(x_k), B^{\mathrm{c}}(x_k)] \end{aligned} \quad \text{Set Difference} \tag{114}$$

$$A \subseteq B = A(x_k) \leqslant B(x_k) \text{ for all } k \quad \text{Subsethood} \tag{115}$$

$$|\mathrm{A}| = \sum_k A(x_k) \quad \text{Cardinality} - \text{Set} \tag{116}$$

$$|A \cap B| = \sum_k \min[A(x_k), B(x_k)] \quad \text{Cardinality} - \text{Set Intersection} \tag{117}$$

$$|A \cup B| = \sum_k \max[A(x_k), B(x_k)] \quad \text{Cardinality} - \text{Set Union} \tag{118}$$

$$|A| = |A - B| + |A \cap B| \quad \text{Cardinality} - \text{Set} \tag{119}$$

$$|A \cup B| = |A| + |B| - |A \cap B| \quad \text{Cardinality} - \text{Set Union} \tag{120}$$

$$|A \cup B| = |A - B| + |B - A| + |A \cap B| \quad \text{Cardinality} - \text{Set Union} \tag{121}$$

Relations, which are also sets, play an important role in set theory and in the similarity theory, but due to space limitations are not formally considered in this work.

Acknowledgments

The authors would like to thank Tom Doman for his constructive comments on the original version of this manuscript, and Mark Johnson, Mic Lajiness, John Van Drie, and Tudor Oprea for helpful discussions. Special thanks are given to Jurgen Bajorath and Jose Medina-Franco, for providing several figures and for their helpful comments.

References

1. Rouvray, D. (1990) The evolution of the concept of molecular similarity. In *Concepts and Applications of Molecular Similarity*, M.A. Johnson and G.M. Maggiora, Eds., Wiley, New York, Chapter 2.
2. Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discovery Today* **7**, 903–911.
3. Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems*. Research Studies Press, Letchworth.
4. Johnson, M.A. and Maggiora, G.M., Eds. (1990) *Concepts and Applications of Molecular Similarity*. Wiley, New York.
5. Dean, P.M., Ed. (1994) *Molecular Similarity in Drug Design*. Chapman & Hall, Glasgow.
6. Tversky, A. (1977) Features of similarity. *Pyschol. Rev.* **84**, 327–352.
7. Chen, X. and Brown, F.K. (2007) Asymmetry of chemical similarity. *Chem. Med. Chem.* **2**, 180–182.
8. Willett, P., Barnard, J.P., and Downs, G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
9. Bender, A. and Glen, R.C. (2004) Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2**, 3204–3218.
10. Johnson, M.A. (1989) A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* **3**, 117–145.
11. Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling*. Springer, New York.
12. Jolliffe, I.T. (2002) *Principal Component Analysis (Second Edition)*. Springer, New York.

13. Domine, D., Devillers, J., Chastrette, M., and Karcher, W. (1993). Non-linear mapping for structure-activity and structure-property modeling. *J. Chemometrics* **7**, 227–242.
14. Rush, J.A. (1999) Cell-based methods for sampling high-dimensional spaces. In *Rational Drug Design*, Truhlar, D.G., Howe, W.J., *et al.*, Eds., Springer, New York, pp. 73–79.
15. Rohrbaugh, R.H. and Jurs, P.C. (1987) Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal. Chim. Acta* **199**, 99–109.
16. Verloop, A. (1987) *The STERIMOL Approach to Drug Design*. Marcel Dekker, New York.
17. Mulliken, R.S. (1955) Electronic population analysis on LCAO-MO molecular wave functions. I. *J. Chem. Phys.* **23**, 1833–1840.
18. Stanton, D.T.; Jurs, P.C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Anal. Chem.* **62**, 2323–2329.
19. Kier, L.B. (1989) An index of molecular flexibility from kappa shape attributes. *Quant. Struct.-Act. Relat.* **8**, 221–224.
20. Kvasnička, V. and Pospíchal, J. (1989) Two metrics for a graph-theoretical model of organic chemistry. *J. Math. Chem.* **3**, 161–191.
21. Kvasnička, V. and Pospíchal, J. (1991) Chemical and reaction metrics for graph-theoretical model of organic chemistry. *J. Mol. Struct. (Theochem.)* **227**, 17–42.
22. Randić, M. (1992) Representation of molecular graphs by basic graphs. *J. Chem. Inf. Comput. Sci.* **32**, 57–69.
23. Baskin, I.I., Skvortsova, M.I., Stankevich, I.V., and Zefirov, N.S. (1995) On the basis of invariants of labeled molecular graphs. *J. Chem. Inf. Comput. Sci.* **35**, 527–531.
24. Skvortsova, M.I., Baskin, I.I., Stankevich, I.V., Palyulin, V.A., and Zefirov, N.S. (1998) Molecular similarity. I. Analytical description of the set of graph similarity measures. *J. Chem. Inf. Comput. Sci.* **38**, 785–790.
25. Ginn, C.M.R., Willett, P., and Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion. *Perspec. Drug Disc. Design* **20**, 1–16.
26. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **44**, 1177–1185.
27. Whittle, M., Gillet, V.J., Willett, P., Alexander, A., and Loesel, J. (2004) Enhancing the effectiveness of virtual screening by fusing nearest-neighbor lists: A comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **44**, 1840–1848.
28. Whittle, M., Gillet, V.J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: Similarity and group fusion. *J. Chem. Inf. Model.* **46**, 2206–2219.
29. Mestres, J., Rohrer, D.C., and Maggiora, G.M. (1999) A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput.-Aided Mol. Design* **13**, 79–93.
30. Trinajstić, N. (1992) *Chemical Graph Theory*. CRC Press, Boca Raton, Florida.
31. Harary, F. (1969) *Graph Theory*. Addison-Wesley Publishing Company, Reading, Massachusetts.
32. Raymond, J.W. and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Design* **16**, 521–533.
33. Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C., and Labaudiniere, R.F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med Chem.* **42**, 3251–3264.
34. Devillers, J. and Balaban, A.T., Eds. (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers, Amsterdam, The Netherlands.
35. Pearlman, R.S. and Smith, K.M. (1998) Novel software tools for chemical diversity. *Perspec. Drug Disc. Design* **9/10/11**, 339–353.
36. Halmos, P.R. (1958) *Finite-Dimensional Vector Spaces, Second Edition*. D. Van Nostrand Company, Inc., Princeton, New Jersey.
37. Mestres, J., Rohrer, D.C., and Maggiora, G.M. (1997) MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* **18**, 934–954.
38. Thorner, D.A., Willett, P., Wright, P.M., and Taylor, R. (1997) Similarity searching in files of three-dimensional chemical structures: Representation and searching of molecular electrostatic potentials using field-graphs. *J. Comput.-Aided Mol. Design* **11**, 163–174.

39. Du, Q., Arteca, G.A., and Mezey, P.G. (1997) Heuristic lipophilicity potential for computer-aided rational drug design. *J. Comput.-Aided Mol. Design* **11**, 503–515.
40. Oden, J.T. and Demkowicz, L.F. (1996) *Applied Functional Analysis.* CRC Press, Boca Raton, Florida.
41. Petke, J.D. (1993) Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem.* **14**, 928–933.
42. Cramer, R.D., Patterson, D.E., and Bunce, J.D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Amer. Chem. Soc.*, **110**, 5959–5967.
43. Bandemer, H. and Näther, W. (1992) *Fuzzy Data Analysis.* Kluwer Academic Publishers, Dordrecht, The Netherlands.
44. Kaufmann, A. and Gupta, M.M. (1985) *An Introduction to Fuzzy Arithmetic – Theory and Applications.* Van Nostrand Reinhold, New York.
45. McGregor, J. and Willett, P. (1981) Use of a maximal common subgraph algorithm in the automatic identification of the ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.* **21**, 137–140.
46. Johnson, M. (1985) Relating metrics, lines, and variables defined on graphs to problems in medicinal chemistry. In *Graph Theory and its Applications to Algorithms and Computer Science*, Y. Alavi *et al.*, Eds., Wiley, New York, pp.457–470.
47. Hagadone, T.R. (1992) Molecular substructure similarity searching: Efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* **32**, 515–521.
48. Rusinko, A., Farmen, M.W., Lambert, C.G., and Young, S.S. (1997) SCAM: Statistical classification of activities of molecules using recursive partitioning. 213th ACS Natl. Meeting, San Francisco, CA, CINF 068.
49. James, C.A., Weininger, D., and Delany, J. (2002) *Daylight Theory Manual.* Daylight Chemical Information Systems, Inc.
50. Kanerva, P. (1990) *Sparse Distributed Memory.* MIT Press, Cambridge, Massachusetts, pp. 26–27.
51. Klir, G.J. and Yuan, B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall PTR, Upper Saddle River, New Jersey.
52. Miyamoto, S. (1990) *Fuzzy Sets in Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers, Dordrecht, The Netherlands.
53. Maggiora, G.M., Petke, J.D., and Mestres, J. (2002) A general analysis of field-based molecular similarity indices. *J. Math. Chem.* **31**, 251–270.
54. Hurst, T. and Heritage, T. (1997) HQSAR – A highly predictive QSAR technique based on molecular holograms. 213th ACS Natl. Meeting, San Francisco, CA, CINF 019.
55. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**, 2894–2896.
56. Xue, L., Godden, J.W., and Bajorath, J. (1999) Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **39**, 881–886.
57. Wikipedia website, http://en.wikipedia.org/wiki/Euclidean_vector (*Last accessed October 22, 2009*).
58. Hyvarinen, A., Karhunen, J., and Oja, E. (2001) *Independent Component Analysis.* Wiley, New York.
59. Kay, D.C. (1988) *Theory and Problems of Tensor Calculus, Schaum's Outline Series.* McGraw-Hill, New York.
60. Hodgkin, E.E. and Richards, W.G. (1987) Molecular similarity based on electrostatic potential and electric fields. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **14**, 105–110.
61. Good, A.C. and Richards, W.G. (1998) Explicit Calculation of 3D molecular similarity. *Perspec. Drug Disc. Design* **9/10/11**, 321–338.
62. Lemmen, C. and Lengauer, T. (2000) Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Design* **14**, 215–232.
63. Güner, O.F., Ed. (2000) *Pharmacophore Perception, Development and Use in Drug Design.* International University Line, La Jolla.
64. Mansfield, M.L., Covell, D.G., and Jernigan, R.L. (2002) A new class of molecular shape descriptors. Theory and properties. *J. Chem. Inf. Comput. Sci.* **42**, 259–273.
65. Grant, J.A., Gallardo, G.A., and Pickup, J.T. (1996) A fast method of molecular shape comparison. A simple application of a Gaussian description of molecular shape. *J. Comp. Chem.* **17**, 1653–1666.
66. Blinn, J.R., Rohrer, D.C., and Maggiora, G.M. (1998) Field-based similarity forcing in energy minimization and molecular matching. In *Pacific Symposium on Biocomputing '99*, R.B. Altman, *et al.*, Eds., World Scientific, Singapore, pp. 415–424.
67. Labute, P. (1999) Flexible alignment of small molecules. *J. Chem. Comput. Group*,

Spring 1999 Edition [http://www.chemcomp.com/feature/malign.htm].
68. Christoffersen, R.E. and Maggiora, G.M. (1969) *Ab initio* calculations on large molecules using molecular fragments. Preliminary investigations. *Chem. Phys. Letts.* **3**, 419–423.
69. Szabo, A. and Ostlund, N.S. (1982) *Modern Quantum Chemistry – Introduction to Advanced Electronic Structure Theory*. Macmillan Publishing Company, New York.
70. Kearsley, S.K. and Smith, G.M. (1990) An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Meth.* **3**, 615–633.
71. Lemmen, C., Hiller, C., and Lengauer, T. (1998) RigFit: A new approach to superimposing ligand molecules. *J. Comput.-Aided Mol. Design* **12**, 491–502.
72. Good, A.C., Hodgkin, E.E., and Richards, W.G. (1992) Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **32**, 188–191.
73. Carbó, R. and Calabuig, B. (1990) Molecular similarity and quantum chemistry. In *Concepts and Applications of Molecular Similarity*, M.A. Johnson and G.M. Maggiora, Eds.,Wiley-Interscience, New York, pp. 147–171.
74. Petitjean, M. (1995) Geometric molecular similarity from volume based distance minimization: Application to Saxitoxin and Tetrodotoxin. *J. Comput. Chem.* **16**, 80–90.
75. Petitjean, M. (1996) Three-dimensional pattern recognition from molecular distance minimization. *J. Chem. Inf. Comput. Sci.* **36**, 1038–1049.
76. Ballester, P.J. and Richards, W.G. (2007) Ultrafast shape recognition for similarity search in molecular databases. *Proc. Roy. Soc.* **A463**, 1307–1321.
77. Nissink, J.W.M., Verdonk, M.L., Kroon, J., Mietzner, T., and Klebe, G. (1997) Superposition of molecules: Electron density fitting by application of Fourier transforms. *J. Comput. Chem.* **18**, 638–645.
78. Keseru, G.M. and Kolossvary, I. (1999) *Molecular Mechanics and Conformational Analysis in Drug Design*. Wiley-Interscience (Blackwell Publishing), New York.
79. Jorgensen, W.L. and Tirado-Rives, J. (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6665–6670.
80. Lee, M.S., Salsbury, F.R., and Olson, M.A. (2004). An efficient hybrid explicit/implicit solvent method for biomolecular simulations. *J. Comput. Chem.* **25**, 1967–1978.
81. Chipot, C. and Pohorille, A., Eds. (2007) *Free Energy Calculations. Theory and Applications in Chemistry and Biology*. Springer, New York.
82. Petit, J., Meurice, N. and Maggiora, G.M. (2009) On the development of a "soft" Rule of Five. *J. Chem. Inf. Model.*, submitted.
83. Stephens, M. A. (1974) EDF Statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**, 730–737.
84. Krishnan, V. (2006) *Probability and Random Processes*. Wiley-Interscience, Hoboken, New Jersey.
85. Martin, Y.C. (2001) Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **3**, 231–250.
86. Seilo, G. (1998) Similarity measures: Is it possible to compare dissimilar structures? *J. Chem. Inf. Comput. Sci.* **38**, 691–701.
87. Medina-Franco, J.L., Martínez-Mayorga, K., Giulianotti. M.A., Houghten, R.A., and Pinilla, C. (2008) Visualization of chemical space in drug discovery. *Curr. Comput.-Aided Drug Design* **4,** 322–333.
88. Oprea, T.I. and Gottfries, J. (2001) Chemography: The art of navigating in chemical space. *J. Comb. Chem.*, **3,** 157–166.
89. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; and Woolsey, J. DrugBank: A comprehensive resource for *in silico* drug discovery and exploration, *Nucl. Acids Res.* **2006**, *34*, D668-D672. (http://www.drugbank.ca/databases. *Accessed July 6, 2009*)
90. Austin, C.P., Brady, L.S., Insel, T.R., and Collins, F.S. (2004) Molecular biology: NIH Molecular libraries initiative. *Science* **306**, 1138–1139. This library is freely accessible by querying 'MLSMR'in PubChem (http://pubchem.ncbi.nlm.nih.gov. *Accessed October 29, 2009*)
91. Patterson, D.E., Cramer, R.D., Ferguson, A. M., Clark, R.D., and Weinberger, L.E. (1996) Neighborhood behavior: A useful concept for validation of molecular diversity. *J. Med. Chem.* **39**, 3049–3059.
92. Bellman, R.E. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton, New Jersey.
93. Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer, New York.
94. Bishop, C. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.
95. Raghavendra, A.S. and Maggiora, G.M. (2007) Molecular basis sets – A general similarity-based approach for representing

chemical spaces. *J. Chem. Info. Model.* **47**, 1328–1340.
96. Simovici, D.A. and Djeraba, C. (2008) *Mathematical Tools for Data Mining: Set Theory, Partial Orders, Combinatorics.* Springer, London, UK.
97. Lee, J.A. and Verleysen, M. (2007) *Nonlinear Dimensionality Reduction.* Springer, New York.
98. Walker, P.D., Maggiora, G.M., Johnson, M. A., Petke, J.D., and Mezey, P.G. (1995) Shape group-analysis of molecular similarity - Shape similarity of 6-membered aromatic ring-systems. *J. Chem. Inf. Comput. Sci.* **35**, 568–578.
99. Rarey, M. and Dixon, J.S. (1998) Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Design* **12**, 471–490.
100. Agrafiotis, D.K. and Lobanov, V.S. (2000) Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **40**, 1356–1362.
101. Rassokhin, D., Lobanov, V.S. and Agrafiotis, D.K. (2000) Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comput. Chem.* **21**, 1–14.
102. Xie, D., Tropsha, A., and Schlick, T. (2000) An efficient projection protocol for chemical databases: Singular value decomposition combined with truncated-Newton minimization. *J. Chem. Inf. Comput. Sci.* **40**, 167–177.
103. Kruskal, J. (1977) The relationship between multidimensional scaling and clustering in *Classification and Clustering.* J. Van Ryzin, Ed., Academic Press, New York.
104. Gower, J.C. (1966) Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338.
105. Diamantaras, K.I. and Kung, S.Y. (1996) *Principal component neural networks – Theory and Applications.* Wiley, New York.
106. Benigni, R. and Giuliani, A. Analysis of distance matrices for studying data structures and separating classes. *Struct.-Act. Relat.* **12**, 397–401.
107. Gower, J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–74.
108. Gower, J.C. (1984) Distance matrices and their Euclidean approximation. In *Data Analysis and Informatics, III*, E. Diday *et al.*, Eds., Elsevier Science Publishers B.V. (North-Holland).
109. Gower, J.C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *J. Classific.* **3**, 5–48.
110. Benigni, R. (1994) EVE, a distance-based approach for discriminating non-linearly separable groups. *Quant. Struct.-Act. Relat.* **13**, 406–411.
111. Tenenbaum, J.B., de Silva, V., and Langford, J.V. (2000) A global geometric framework for non-linear dimensionality reduction. *Science* **290**, 2319–2323.
112. Roweis, S.T. and Saul, L.K. (2000) Nonlinear dimensionality reduction by local linear embedding. *Science* **290**, 2323–2326.
113. Friedman, J. and Tukey, J. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C23**, 881–889.
114. Agrafiotis, D.K. (2003) Stochastic proximity embedding. *J. Comput. Chem.* **24**, 1215–1221.
115. Agrafiotis, D.K. and Xu, H. (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **43**, 475–484.
116. Donoho, D.L. and Grimes, C. (2003) Hessian eigenmaps: Local linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci U. S. A.* **100**, 5591–55.
117. Maggiora, G.M., Shanmugasundaram, V., Lajiness, M.S., Doman, T.N., and Schulz, M.W. (2005) A practical strategy for directed compound acquisition. In *Chemoinformatics in Drug Discovery*, T.I. Oprea, Ed., pp. 317–332.
118. Maggiora, G.M. (2006) On outliers and activity cliffs – Why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535 (Editorial).
119. Doweyko, A.M. (2008) QSAR: dead or alive? *J. Comput.-Aided Mol. Design* **22**, 81–89.
120. Johnson, S. (2008) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **48**, 25–26.
121. Guha, R. and Van Drie, J.H. (2008) Assessing how well a modeling protocol capture a structure-activity landscape. *J. Chem. Inf. Model.* **48**, 1716–1728.
122. Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M.S., and Van Drie, J.H. (2009) Navigating structure-activity landscapes. *Drug Disc. Today* **14**, 698–705.
123. Shanmugasundaram, V. and Maggiora, G.M. (2001) Characterizing property and activity landscapes using an information-theoretic approach. *222nd American Chemical Society Meeting*, Division of Chemical Information Abstract no. 77.
124. Renner, S. and Schneider, G. (2005) Scaffold-hopping potential of ligand-based similarity concepts. *Chem. Med. Chem.* **1**, 181–185.

125. Schneider, G., Schneider, P., and Renner, S. (2006) Scaffold hopping: How far can you jump? *QSAR Combin. Sci.* **25**, 1162–1171.
126. Maggiora, G.M. and Shanmugasundaram, V. (2005) An information-theoretic characterization of partitioned property spaces. *J. Math. Chem.* **38,** 1–20.
127. Medina-Franco, J.L., Maggiora, G.M., Giulianotti, M.A., Pinilla, C., and Houghten, R.A. (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Design* **70**, 393–412.
128. Guha, R. and Van Drie, J.H. (2008) Structure-activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646–658.
129. Peltason, L. and Bajorath, J. (2007) SAR index: Quantifying the nature of structure-activity relationships. J. Med. Chem. **50**, 5571–5578.
130. Wawer, M., Peltason, L., Weskamp, N., Teckentrup, A., and Bajorath, J. (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **51**, 6075–6084.
131. Medina-Franco, J.L., Martínez-Mayorga, K., Bender, A., Marín, R.M., Giulianotti, M.A., Pinilla, C., and Houghten, R.A. (2009) Characterization of activity landscapes using 2D and 3D similarity methods: *Consensus activity cliffs. J. Chem. Inf. Model.* **49**, 477–491.
132. Christoffersen, R.E. (1989) *Basic Principles and Techniques of Molecular Quantum Mechanics.* Springer, New York.
133. Schölkopf, B. and Smola, A. (2002) *Learning with Kernels.* MIT Press, Cambridge, MA.
134. Herbrich, R. (2002) *Learning Kernel Classifiers.* MIT Press, Cambridge, MA.
135. Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, UK.
136. Löwdin, P.O. (1992) On linear algebra, the least square method, and the search for linear relations by regression analysis in quantum chemistry and other sciences. *Adv. Quantum Chem.* **23**, 83–126.
137. Meyer, C.D. (2000) *Matrix Analysis and Applied Linear Algebra.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
138. Carlson, B.C. and Keller, J.M. (1957) Orthogonalization procedures and the localization of Wannier functions. *Phys. Rev.* **105**, 102–103.
139. Agrafiotis, D.K., Rassokhin, D.N., and Lobanov, V.S. (2001) Multi-dimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **22**, 1–13.
140. Kauvar, L.M., Higgins, D.L., and Villar, H.O., *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **2**, 107–118.
141. Randic, M. (1991) Resolution of ambiguities in structure-property studies by use of orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **31**, 311–320.
142. Randic, M. (1991) Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct.(Theochem)* **233**, 45–59.
143. Randic, M. (1993) Fitting non-linear regressions by orthogonalized power series. *J. Comput. Chem.* **14**, 363–370.

Chapter 3

The Ups and Downs of Structure–Activity Landscapes

Rajarshi Guha

Abstract

In this chapter we discuss the landscape view of structure–activity relationships (SARs). The motivation for such a view is that SARs come in a variety of forms, such as those where small changes in structure lead to small changes in activity or where small structural lead to significant changes in activity (also termed activity cliffs). Thus, an SAR dataset is viewed as a landscape comprised of smooth plains, rolling hills, and jagged gorges. We review the history of this view and early quantitative approaches that attempted to encode the landscape. We then discuss some recent developments that directly characterize structure–activity landscapes, in one case with the goal of highlighting activity cliffs while the other allows one to resolve different *types* of SAR that may be present in a dataset. We highlight some applications of these approaches, such as predictive model development and SAR elucidation, to SAR datasets obtained from the literature. Finally, we conclude with a summary of the landscape approach and why it provides an intuitive and rigorous alternative to standard views of structure–activity data.

Key words: QSAR, Glucocorticoid, Melanocortin, Activity cliff

1. Introduction

The analysis of structure–activity relationships (SAR) is one of the fundamental tasks in medicinal chemistry. A variety of methods have been described for such analyses, ranging from simple visual inspection of compounds and their activities, scaffold analysis to more quantitative methods such as Quantitative Structure–Activity Relationship (QSAR) models.

The fundamental idea underlying all these approaches is that one somehow systematically identifies structural differences between molecules that lead to differences in their activities. In QSAR methods, this is achieved by correlating numerical descriptions of structural features to the activities of the molecules. A simpler, more visual procedure is to analyze the molecules in a pairwise fashion and highlight structural differences between pairs

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_3, © Springer Science+Business Media, LLC 2011

of molecules and the corresponding differences in their activities. The simplest form is to generate R-group tables, though this can become cumbersome when there are multiple R-groups present in a series of molecules. While these approaches are useful for small collections, one would like to extend this approach to a more automated system that could handle much larger collections, such as those from high-throughput screens. Thus various computational schemes have been developed that try to capture the structure–activity information encoded in pairs of molecules. An early approach is the use of "neighborhood behavior" described by Patterson et al. [1], which simply plotted the differences in biological activity versus the differences in some descriptor for pairs of molecules. While useful, the method does not really provide insight into SAR trends, which is not surprising since it was designed for diversity analysis of libraries. A more recent example of pairwise analyses is the idea of matched molecular pairs [2]. In that work, Leach et al. note that traditionally, identifying pairs of molecules to highlight how certain structural changes lead to improvements (or degradation) in activity has usually been a manual and subjective process. They described a statistical method that provides a probability that indicates that a certain structural change will cause a change in some property (such as aqueous solubility) in a specific direction. Another approach, similar in intent, focuses on identifying molecular transformations to characterize local QSARs [3]. This approach consists of two methods, one for analyzing molecular pairs and deriving the structural transformations that capture the differences in their activities and another method that can transform a query molecule into other related structures based on transformations observed in a previously studied collection.

In all the approaches noted above, there is an implicit assumption that as one moves from one molecule to another, via a structural change, one observes a change in activity. This suggests that one could view a collection of molecules and their activities as a "landscape." For simplicity, if we consider a 3D landscape, then the *x*- and *y*-axes would be defined by some structural descriptor (or combinations of more than two descriptors) and the *z* axis would represent the activity. An example is shown in Fig. 1. The surface is generated by evaluating a set of molecular descriptors for a collection of 81 inhibitors [4, 5] of the melanocortin-4 receptor and then performing a principal components analysis. The first two principal components were plotted on the *x*- and *y*-axes, and a scaled form of their IC_{50}'s on the vertical *z*-axis (activity increasing in the upwards direction). Note that the actual surface is a quadratic interpolation of the 81 points.

As expected, the surface is not flat. More importantly, the peaks and troughs of the surface highlight interesting aspects of the SAR. More specifically, if one considers the peak, it represents a molecule with very good activity. But, just in front, we see a

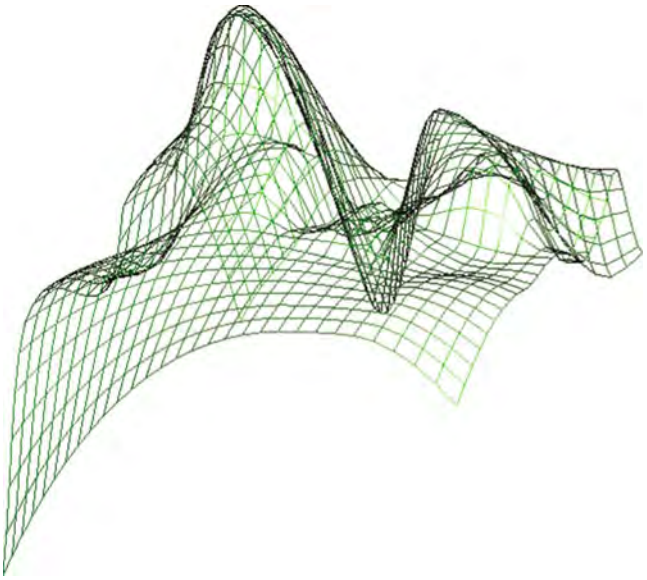


Fig. 1. A schematic representation of a SAR landscape for a set of 81 inhibitors of the melanocortin receptor.

relatively low portion of the surface corresponding to molecules with poorer activity. The key thing to note is that these two sets of molecules are close to each other on *x*- and *y*-axes, i.e., the descriptor space. Hence they are structurally similar, but show a significant difference in activity. Such cases have been termed "activity cliffs" by Maggiora [6] and could be considered as the most interesting parts of an SAR. This view also highlights why QSAR models, especially those based on machine learning, do not always perform very well. These models attempt to encode the landscape. However, the presence of activity cliffs is problematic since they represent discontinuities – which by definition cannot be encoded reliably by a machine learning model.

The landscape view of the SARs for a collection of molecules provides an intuitive way to visualize the relationship between the changes in their structures and the resulting differences in their activities. However, it is also useful to be able to quantify the landscape. In other words, given the landscape view, can we develop metrics that will allow us to quantify the roughness (or smoothness) of the surface. Note that the SAR landscape will, in general, be a high-dimensional hypersurface. As a result, while the 3D visualization of such surfaces is intuitive, methods to characterize such representations should take the high dimensionality into account.

1.1. Outline

The remainder of this chapter discusses recent developments in the characterization of structure–activity landscapes. Subsection 2

presents the formal definitions for two recently described functions aimed at characterizing SARs and activity cliffs. Subsection 3 describes some applications and features of these approaches. Finally, Subsection 4 summarizes the current state of the art.

2. Definitions

SAR landscapes can be characterized in a variety of quantitative ways. Which way is preferred is problem dependent. Thus, if one is interested in directly identifying activity cliffs, one could employ the Structure–Activity Landscape Index (SALI). On the other hand, if one is more interested in characterizing SAR trends, the Structure–Activity Relationship Index (SARI) may be preferably. While the focus of this chapter is the description of SALI, we briefly cover the definitions of both these indices.

2.1. Structure–Activity Landscape Index

The SALI [7] is defined as

$$SAL\ I_{i,j} = \frac{|A_i - A_j|}{1 - sim(i,j)} \quad (1)$$

where A_i and A_j are the activities of molecules i and j and $sim(i,j)$ is the structural similarity between the two molecules. While any form of structural similarity measure can be used for the denominator, most work has employed a Tanimoto similarity coefficient between two binary fingerprints. Of course, for identical molecules (and also those pairs of molecules whose fingerprint representations are identical), the above measure is undefined. For a collection of molecules one can evaluate a symmetric "SALI matrix," in which elements with very large values correspond to activity cliffs. A useful visual summary of the SALI values is to simply plot the matrix as an image, as shown in Fig. 2. In this plot, the axes correspond to molecules and have been in order such that the more active molecules are located to the right and top. As a result, white blocks immediately identify pairs of molecules making up activity cliffs. One can further prioritize such cases by noting that white elements in the top right correspond to an active molecule, which by a small structural change becomes another active molecule. While it is nice to see that the change does not decrease activity, it is not as beneficial as having an *inactive* molecule become an active molecule by a small structural change, as represented by white elements in the bottom right.

While the matrix visualization is a useful summary, it is rather limited and static in nature. We have previously described an alternative visualization of the SALI matrix that lends itself to a more dynamic view of the SALI data, and as will be described in

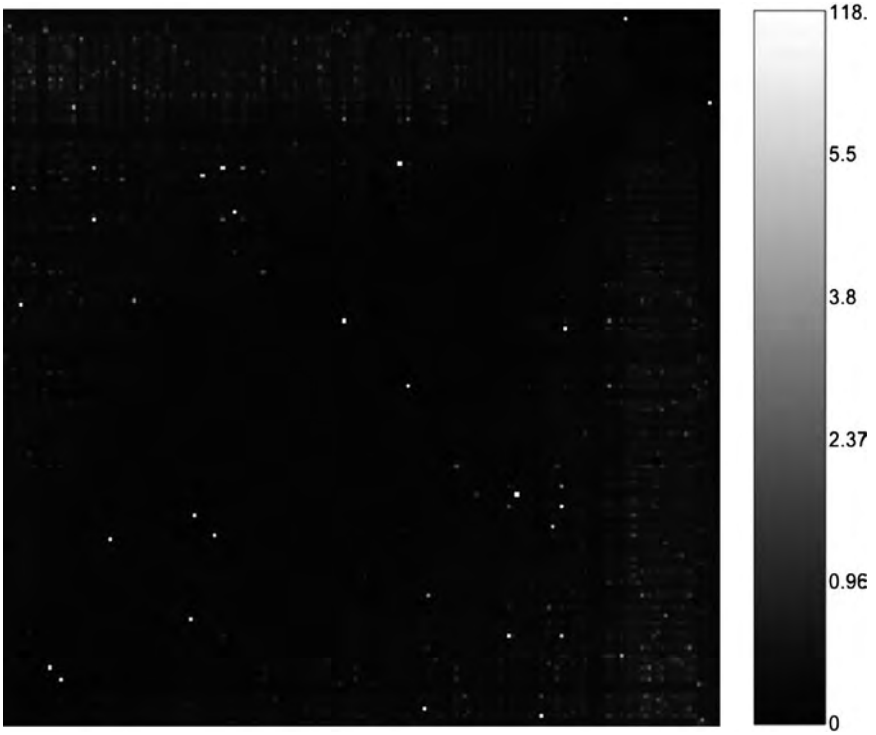


Fig. 2. A visual summary of the SALI values for a collection of 179 molecules. *Lighter colors* represent increasingly significant activity cliffs.

the following sections, presents a powerful tool to characterize SAR datasets and models.

As described in ref. [7], if we consider each molecule to be a node, we can construct a graph such that an edge is drawn between molecules i and j if $SALI_{i,j} > C$ where C is some arbitrary cutoff. Clearly, with smaller values of C, more and more molecules will be connected leading to a relatively complex network. However, at high values of C, only those molecular pairs that have very high SALI values will remain in the network. In other words, one can use a high value of C to quickly "zoom in" on the most significant activity cliffs. An example of such a SALI network is shown in Fig. 3. The network is laid out such that the edges are directed with the molecule at the tail having poorer activity than the molecule at the head. In addition, as one goes from top to bottom of the graph, the activity improves. It should be noted that even though some nodes are at the same horizontal level, they do not have identical activities. This aspect is an artifact of the layout algorithm.

While the SALI network is a useful way to explore series of structural changes that lead to improvements in activity (especially with the interactive tool described in ref. [7]), it also allows us to integrate a variety of external information. A simple example is to

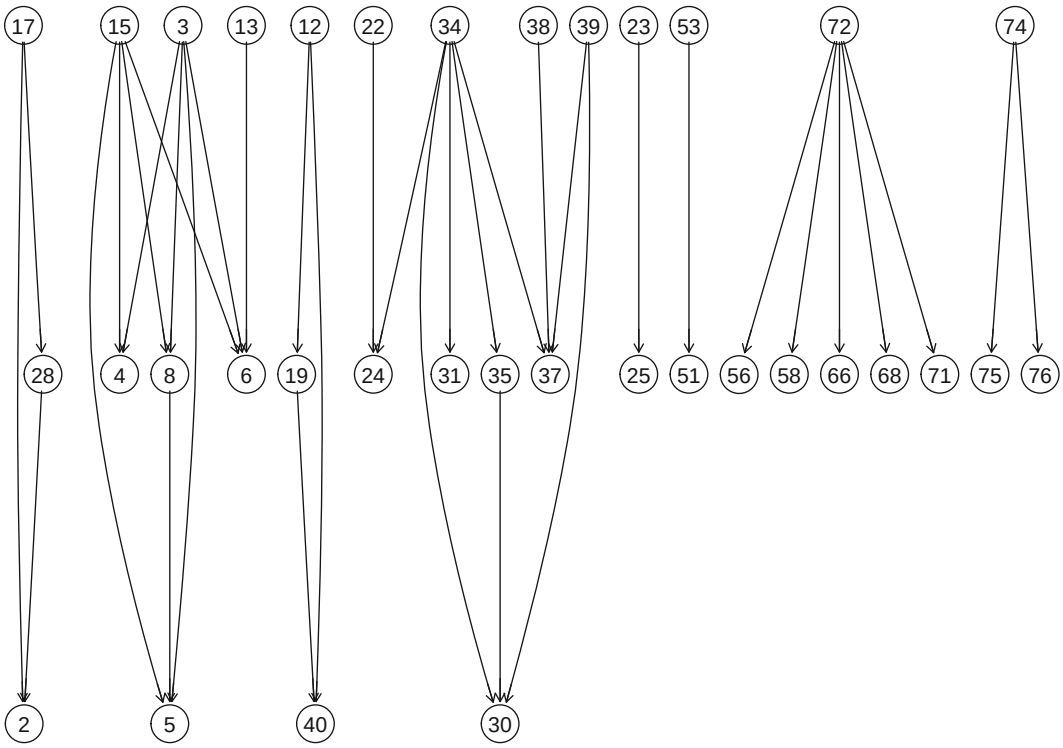


Fig. 3. An example of a SALI network.

overlay synthetic feasibility in going from one molecule to another. Thus for every pair of molecules, one can evaluate a synthetic feasibility score (SFS) that characterizes the ease with which a molecule can be synthesized from another. Since most SFSs consider individual molecules [8, 9], a trivial way to consider pairs of molecules is to simply take the difference of the SFSs of the two molecules in question. This value can then be used to color code the edges of the SALI network.

While the definition above employs IC_{50}'s, one can replace it with any value that tracks activity. Thus, one could employ docking scores in the numerator of Eq. 3.1, the assumption being that docking scores correlate to binding affinity (though this is not always the case). Rather than employing surrogates for activity, one could also consider employing entire dose–response curves in place of single point activities, as described in Subsection 3.1.

2.2. Structure–Activity Relationship Index

An alternative approach and more general approach to characterizing SARs in a dataset was described by Peltason and Bajorath [10] in which they defined the SAR Index (SARI). The approach is based on the categorization of SARs into continuous, discontinuous, and heterogeneous, such that the SARI quantifies these categories when applied to a given dataset and target. In their terminology, a discontinuous SAR would correspond to the

presence of activity cliffs whereas a continuous SAR would correspond to smooth regions of the SAR landscape (where a small change in structure leads to little or no change in activity). Mathematically, the SARI is defined as

$$SARI = \frac{1}{2}(score_{cont} + 1 - score_{disc}) \quad (2)$$

where $score_{cont}$ (the continuity score) measures the potency weighted structural diversity (based on pairwise ligand similarity calculations) of a group of compounds and $score_{disc}$ (the discontinuity score) measures the average potency difference between pairs of compounds. It is important to note that the SARI value is defined on a *set* of compounds, rather than individual pairs (though the latter are employed in the calculation of the SARI value for a collection). In general, the SARI values will range from 0 to 1, where higher values correspond to more continuous SARs and smaller values to discontinuous SARs (i.e., activity cliffs). The detailed mathematical derivations of these scores can be found in the original work. The authors then apply this methodology to a variety of compounds classes targeting a variety of receptors. The results indicate that the SARI values are able to characterize the SARs present in each of these datasets, highlighting smooth as well as discontinuous portions of the individual SAR landscapes. One of the drawbacks of this approach is that, it is based on prespecified cutoff values. For example, the discontinuity score employs the Tanimoto similarity value between pairs of molecules, but only for those pairs whose T_c crosses a threshold of 0.6. While they report that increasing the threshold value did not significantly change the final results, this is still an arbitrary constant. An example, taken from the original work is shown in Fig. 4. The four structures are taken from a collection of thromboxane synthase inhibitors. The numbers on the edges indicate the Tanimoto similarity between the structures. As reported, the SARI value for the set of 23 compounds was 0.46, which is due to a mix of continuous and discontinuous SARs present in this dataset (highlighted in Fig. 4). Due to the presence of both types of SARs in a single dataset, the authors characterize this as a *heterogeneous relaxed* SAR.

3. Case Studies

3.1. Richer Activity Data

Equation 3.1 above employs single point activities. While informative, it is a one-dimensional view of the activity of a molecule. Specifically, it describes the potency but does not provide any information regarding the efficacy of the molecule. This information can be obtained from dose–response assays. In such cases, one

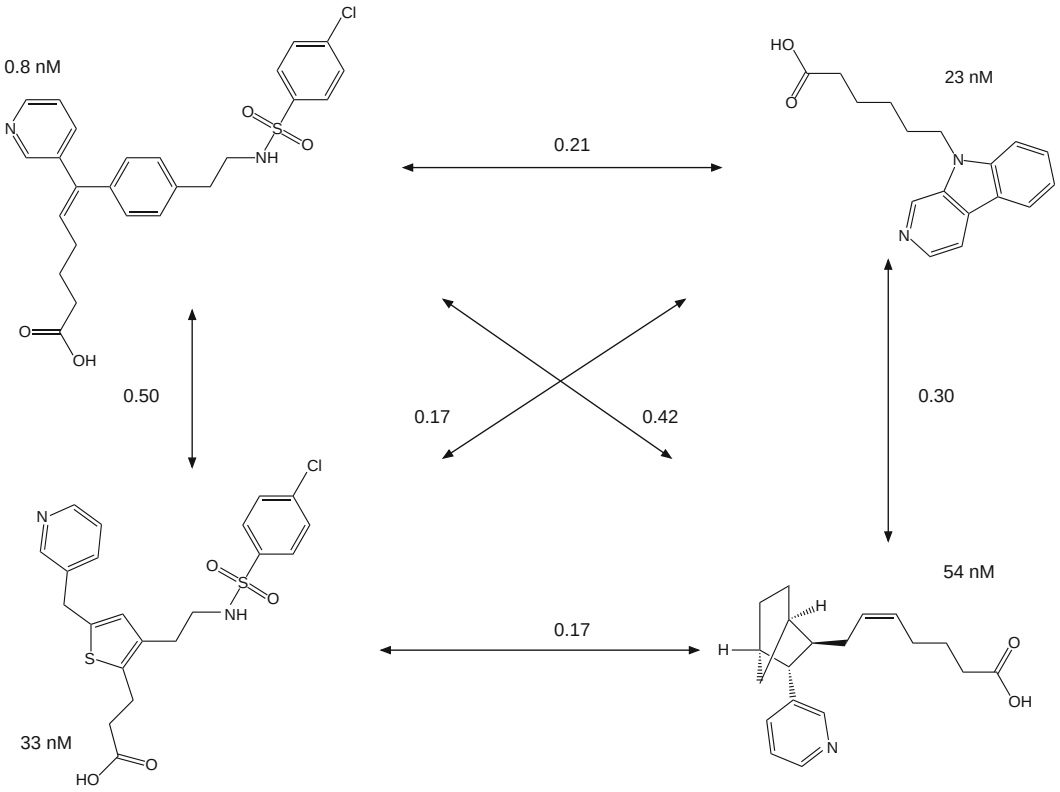


Fig. 4. A set of molecules highlighting the continuous and discontinuous SAR trends characterized by SARI. The edge labels indicate Tanimoto similarity between the corresponding molecules. Taken from Peltason and Bajorath [10].

can obtain a dose–response curve, which is generated by fitting the Hill equation [11].

$$y = S_0 + \frac{S_{\text{inf}} - S_0}{1 - 10^{(-\log AC_{50} - x)n}} \tag{3}$$

One now has four parameters describing the biological activity of the compound: S_0, S_{inf}, n, and $\log AC_{50}$. Figure 5 shows a dose–response curve fit using the Hill equation, highlighting the meaning of the four Hill parameters. S_0 and S_{inf} represent the activities at the initial and final portions of the curve. The AC_{50} is the concentration at which one observes 50% activity. Note that the curve fit provides us with $\log AC_{50}$ from which we calculate the AC_{50}. Finally, the slope of the curve between initial and final plateau is denoted by n.

As a result, rather than simply using the activity in the numerator of Eq. 3.1, one could employ the four Hill parameters. Thus, Eq. 3.1 could be rewritten as

$$SALI_{i,j} = \frac{D(\{P_i\} - \{P_j\})}{1 - sim(i,j)} \tag{4}$$

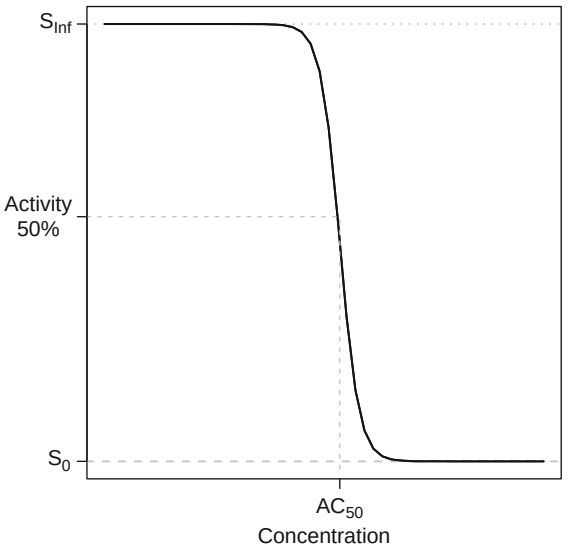


Fig. 5. A schematic diagram of a dose–response curve fit using the Hill equation. The plot is annotated with three of the four Hill parameters. The remaining parameter is the slope of the curve between the two exponential portions of the curve.

where $D()$represents the Euclidean distance function and P_i and P_j are the Hill curve parameters for molecules I and j. When the SALI matrix is regenerated using Eq. 3.4, we get very similar results to that obtained using Eq. 3.1. While the most significant cliffs (i.e., largest values in the SALI matrix) are identical in both cases, the use of curve parameters leads to some differences in the less significant cliffs. Compared to SALI matrix obtained using AC_{50} values only, a number of compound pairs are identified as cliffs.

Figure 6 shows the SALI matrices derived from these two approaches for a set of 96 compounds tested in an ERK phosphorylation assay.

3.2. SAR Landscapes and Predictive Models

While the use of SALI networks to explore landscapes is useful, one can employ them for other purposes. Guha and Van Drie [12] described the use of SALI networks as a means of measuring the quality of QSAR models.

The key to this application is to realize that the SALI network is directed. Each pair of connected nodes is thus ordered. Given a predictive model and a SALI network at a given cutoff, one can use the predicted values from the model and identify how many of the edges in the network are correctly ordered. As noted above, for large values of the cutoff the network highlights the most significant activity cliffs. Thus one can characterize how well a model predicts the most significant activity cliffs. But this is only a partial view of the predictive ability of the model. Instead, we

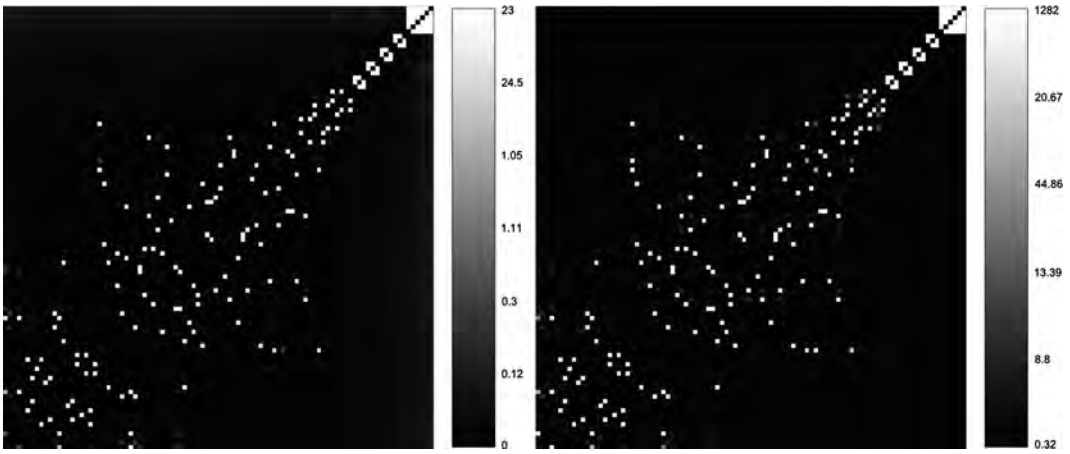


Fig. 6. SALI matrices obtained using simple AC_{50} values (*left*) and dose–response curve parameters for a set of 96 compounds tested for activity in an ERK phosphorylation assay.

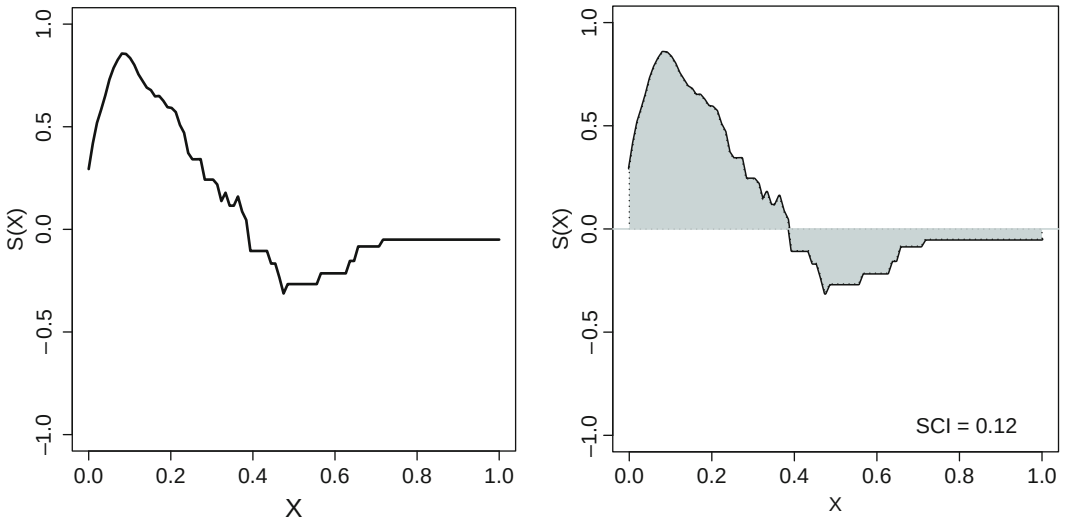


Fig. 7. A SALI curve and characterizing a models' ability to encode an SAR landscape and the SALI Curve Integral (SCI). In both plots, X is the SALI cutoff as a fraction of the maximum SALI value in the dataset and S(X) represents the normalized edge count.

perform this operation on SALI networks generated at a series of cutoffs. The original work evaluated cutoffs from 0 to 100% of the maximum SALI value for the dataset. The result of this is that one can plot the number of correctly predicted edges versus the cutoff. An example of such a plot is shown on the left in Fig. 7, where X represents a fraction of the maximum SALI value for the dataset, and $S(X)$ represents a normalized version of the edge count, such that $S(X) = 0$ indicates that half of the edges are predicted correctly, $S(X) = 1$indicates all the edges are predicted correctly, and $S(X) = -1$ indicates that all the edges are mispredicted. The

curve has two important parts – the region around $S(X) = 0$ and the region around $S(X) = 1$. At $S(X) = 0$ all pairs of molecules are connected and thus the number of edges correctly predicted at this point, is essentially an ordering of the entire dataset. That is, the value of $S(0)$ represents the ability of the model to capture the landscape as a whole. In this sense, $S(0)$ is conceptually equivalent to the root mean square error (RMSE) of the model, and indeed, in simulations, the correlation between these two values is greater than 0.9. The value of $S(1)$, on the other hand, represents the ability of the model to capture the most significant cliffs. In general, while we expect that a good model will capture a large portion of the landscape, we do not expect that machine learning models will have very high values of $S(1)$, since the most significant cliffs represent discontinuities. On the other hand, alternative forms of QSAR models such as docking or pharmacophore methods should be able to accurately identify such significant cliffs, and indeed, SALI curves for such models have been shown to exhibit relatively high values of $S(1)$ [12].

While a useful visual depiction of an ability of the model to encode an SAR landscape, it can be tedious to inspect curves for many molecules. A simple numerical characterization is to represent a SALI curve by the area enclosed between the curve and the $S = 0$ line, as shown on the right in Fig. 7. This metric is termed the SALI Curve Integral (SCI). This approach allows one to summarize the quality (in terms of the ability to encode the landscape) for multiple models in a similar manner. More importantly, the use of the SALI framework to characterize model quality allows one to compare disparate models, on an equal footing. While one could simply use a correlation coefficient to compare the behavior of a docking model and a linear regression model, this is a relatively simplistic measure of quality (and not very robust). On the other hand, the use of the SCI, while still a single number, implicitly contains more information on how well the model is able to characterize the landscape.

3.3. SALI Based Model Selection

While the use of the SCI is an efficient way to explicitly characterize a model's ability to encode the landscape, it also suggests itself as a way to *choose* models that will better encode structure–activity landscapes. In other words, can the SCI (or some other numerical characterization of the SALI curve) be used as a metric to select a set of features that will lead to a model that can capture the details of an SAR landscape with high-fidelity. This is essentially the feature selection problem [13]. A multitude of approaches have been applied to feature selection in QSAR modeling ranging from traditional stepwise approaches to methods such as genetic algorithms, simulated annealing, and particle swarm methods. From the point of view of SAR landscapes, it does not really matter which specific method is used. Rather, for any given method,

we use the SALI curve as the objective function that is being optimized.

As an example, we considered the 81 molecule melanocortin-4 inhibitor dataset [4, 5]. Using a set of topological descriptors, we performed feature selection, using a genetic algorithm, to identify a set of good linear regression models. Initially, we considered an objective function that simply minimized the root mean square error (RMSE). Then we replaced the objective function with one based on the SALI curve. Specifically, we defined the function as

$$S(0) + \frac{1}{2}(S_{\max} - S_{\min}) \tag{5}$$

where $S(0)$is the value of the SALI curve at $x = 0$ (i.e., the portion of the curve that is analogous to the RMSE for the whole dataset), $S_{\max}$ and $S_{\min}$ are the maximum and minimum values of the SALI curve. This approach was chosen based on the fact that while maximizing the plateau region of the SALI curve would lead to models capturing more of the significant activity cliffs, it would also likely lead to overfitting – since the most significant cliffs would correspond to discontinuities, which would have to be "memorized" by the model. Using these two objective functions, we then identified a set of 3- and 6-descriptor models. Figure 8 displays a summary of the RMSEs of the models obtained from the two approaches. While it is clear that the SALI based approaches lead to slightly poorer overall RMSEs, the resultant models do have better SCI values compared to the models obtained purely based on optimizing the RMSE. Thus for the 3-descriptor models obtained by optimizing RMSE, the SCI values range from 0.004 to 0.17 whereas for the SALI derived

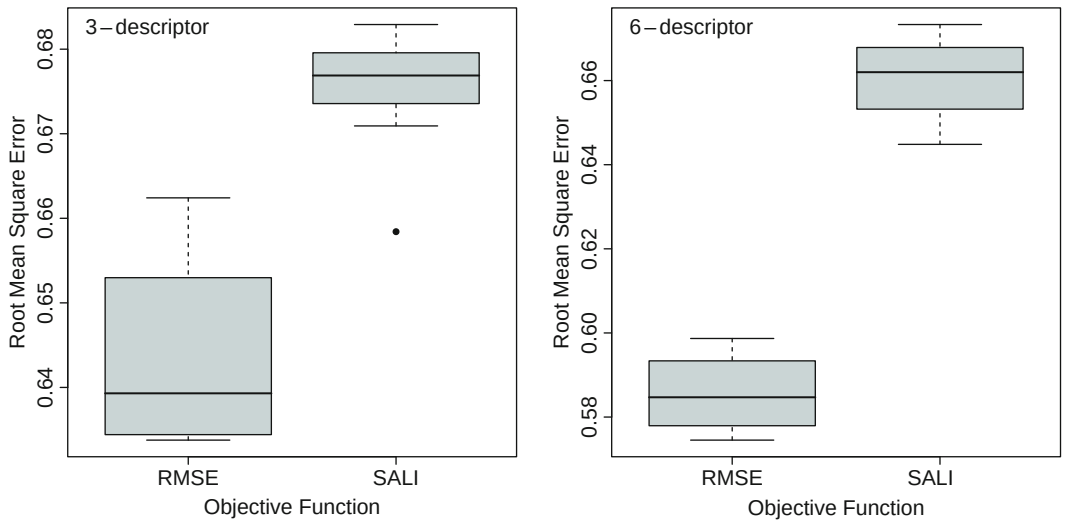


Fig. 8. A summary of the RMSEs of linear regression models obtained using traditional RMSE-based feature selection and SALI curve-based feature selection.

models, the SCI values range from 0.31 to 0.46. While not very good values, this is not surprising, as the models themselves do not exhibit significant predictive ability. For the 6-descriptor RMSE derived models the SCIs range from 0.0 to 0.16 whereas for the SALI derived models we observed SCIs from 0.41 to 0.58. Clearly, feature selection based on the SALI curves does lead to models with better ability to encode the SAR landscape. But this comes at a cost. By requiring that models encode at least some of the significant activity cliffs, this introduces some degree of overfitting. This downside could trivially be balanced by modifying the objective function to take into account possible overfitting (such as introducing a penalty for poor cross-validation statistics).

3.4. Network Analysis of SAR Landscapes

As described previously, both SALI and SARI focus on a network representation of a SAR dataset. In the case of SALI, the network view can be thought of as a transformation of the data that allows subsequent analyses. In other words, it does not necessarily encode physical features of the SAR(s) in the dataset. Thus, while one could evaluate a variety of network metrics such as vertex degree distribution, centrality, and so on, such values likely do not have physical relevance. Even so, it is interesting to investigate whether a numerical analysis of the structure of a SALI network can allow us to indicate that one or more SARs are present or not. As an example, we consider a SALI network constructed from a set of 62 glucocorticoid [14, 15] inhibitors. The original network is shown in Fig. 9 at a cutoff of 30% of the maximum SALI value in the dataset. Clearly, the actual SALI network has a number of disconnected components. Furthermore, the graph appears to exhibit a nonrandom structure. In other words, the connectivity in the graph appears to be different from what would be observed in a random graph (i.e., the same number of nodes but connected randomly). To investigate this observation, we considered random graphs with the same number of nodes as in the original SALI graph, and edges generated according to Erdos–Renyi model [16]. We then evaluated various metrics such as the degree distribution, closeness [17], and transitivity [18] values. Figure 10 summarizes the degree distributions of the original SALI graph and the 100 random graphs. In the latter case, the degree distribution is obtained by taking the mean

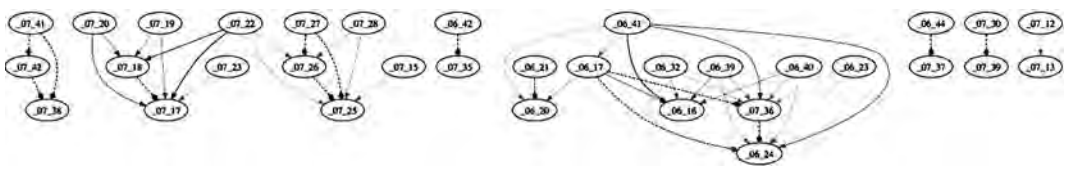


Fig. 9. A SALI graph for a set of 62 inhibitors of the glucocorticoid receptor generated at a SALI cutoff of 30% of the maximum SALI value in the dataset.

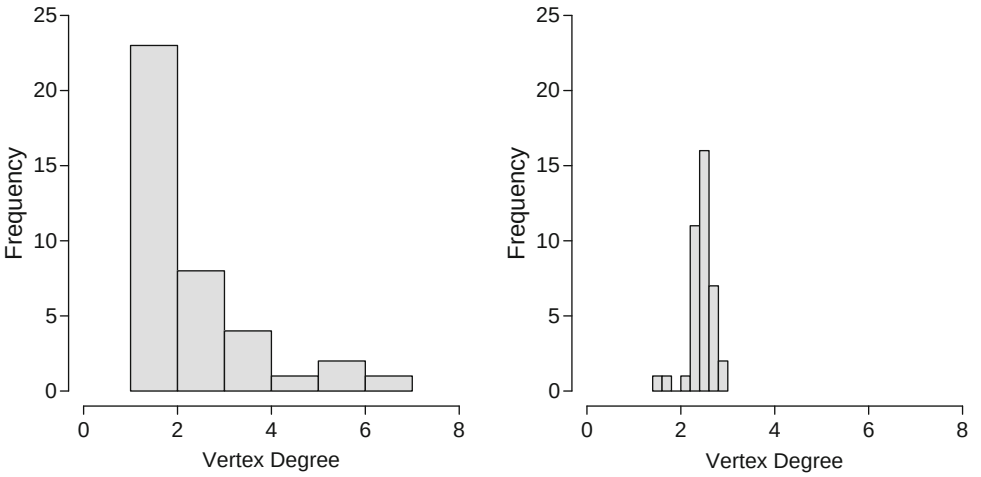


Fig. 10. A comparison of vertex degree distributions for the original SALI graph (*left*) and 100 random graphs (*right*), with the same number of nodes as in the original SALI graph, but with edges determined by the Erdos–Renyi model.

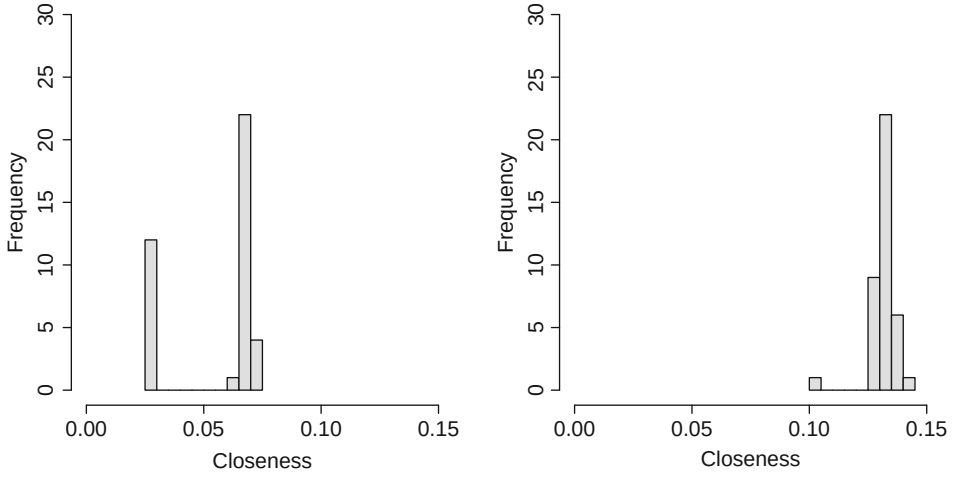


Fig. 11. A comparison of vertex closeness distributions for the original SALI graph (*left*) and 100 random graphs (*right*), with the same number of nodes as in the original SALI graph, but with edges determined by the Erdos–Renyi model.

degree of each node across the 100 random graphs. Clearly, the random graphs show a much lower degree of variation in vertex degree compared to the original SALI network. Similarly, one can compare the closeness values. As with the vertex degrees, this is also a distribution, and Fig. 11 compares the closeness for the original SALI graph and the same 100 random graphs. It is clear that the network metrics calculated for the SALI graphs are distinct from those derived from random graphs. This observation suggests that one could employ random graphs to determine a p-value for a given SALI graph, using it as an indication that the structure of a given SALI graph is not due to chance effects.

The SARI approach has been employed to annotate and analyse network-like similarity graphs [19], where the nodes are chemical structures and nodes are connected by an edge if their Tanimoto similarity is greater than some cutoff. In a study by Wawer et al. [19], the SARI values were used to scale the size of nodes to highlight the discontinuity of the SAR in those regions of the graph. These SARI encoded network-like similarity graphs have been used to characterize different aspects of a chemical dataset, ranging from pairwise similarity relationships (via the edges) to SAR discontinuity and activity cliffs (via the SARI based scaling of node sizes). While the authors described a variety of control experiments as well as an analysis of the topology of these network-like similarity graphs in terms of clustering, it is clear that these graphs also exhibit nonrandom structures. Thus, one might expect that a similar approach to that described above for SALI graphs could be profitably applied to these networks to derive a measure of significance.

4. Conclusions

While the concepts underlying SAR landscapes have been employed by practicing medicinal chemists in a multitude of real world projects, the ability to quantitatively characterize these principles opens up a number of useful possibilities.

From the standpoint of exploring SAR trends in collections of molecules, the network approach has been shown to easily identify specific instances of SAR series that consistently improve activity [20]. More importantly, the approach allows rapid visual identification of such series amongst large collections such as high throughput screening (HTS) datasets, where the majority of compounds will have little or no activity. The network paradigm is also a useful approach to obtain rapid visual summaries of synthetic campaigns, allowing users to highlight changes which lead to improvements in activity [7]. In this sense, the SAR landscape view allows one to easily recapitulate useful synthetic changes, effectively guiding future work on those series.

However, the SAR landscape view is not restricted to visual summaries of SAR datasets. As has been described, the SALI network view can be applied to quantitative models that attempt to encode the SAR landscape. Specifically, one can view approaches such as SALI curves and the SCI as alternatives to RMSE and R^2 that allow a more direct characterization of a predictive model (which might be a machine learning model or a physical approach such as docking and pharmacophore models) in terms of their ability to capture details of the landscape. This view also highlights the problems associated with the different types of

model approaches. Thus, machine learning based models will, in general, never be able to predict significant activity cliffs with great fidelity, simply because they represent discontinuities in the landscape. From a physical point of view, such discontinuities indicate that some specific structural feature (say a specific hydrogen bonding donor) is responsible for the jump in activity. Unless one happens to identify a molecular descriptor that specifically characterizes this physical feature, traditional machine learning models will not effectively encode the behavior of such molecules. Moreover, even if one were to include such a specific descriptor, it would exist primarily to capture the few cases of significant activity cliffs and not contribute much to the other patterns present in the dataset. This would explain why traditional feature selection approaches may not identify such descriptors. On the other hand, SALI based feature selection can identify such features, but at the same time introduces a degree of overfitting into the model. But given that activity cliffs are indicative of specific interactions between ligand and receptor, approaches such as docking and pharmacophore models are inherently better at capturing the details of the landscape. As shown by Guha and Van Drie [12], this is reflected in the SCI values obtained for these types of models. Even in this case, one could profitably employ SCI values to choose between different scoring functions or even explain why specific scoring functions do better than others for a given target. But more generally, the use of landscapes in characterizing quantitative models allows us to compare models built on disparate methodologies on an equal footing.

In conclusion, while the ideas underlying SAR landscapes are not new, the recent developments in quantitative approaches to characterizing such landscapes and applications of these methods to a diverse set of problems ranging from SAR identification to model quality characterization highlight the utility and flexibility of the landscape paradigm.

References

1. Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E., Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996,** *39*, 3049–3059.
2. Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B., Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006,** *49*, 6672–6682.
3. Sheridan, R. P.; Hunt, P.; Culberson, J. C., Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006,** *46* (1), 180–192.
4. Tran, J. A.; Chen, C. W.; Jiang, W.; Tucci, F. C.; Fleck, B. A.; Marinkovic, D.; Arellano, M.; Chen, C.; Tran, J. A.; Chen, C. W.; Jiang, W.; Tucci, F. C.; Fleck, B. A.; Marinkovic, D.; Arellano, M.; Chen, C., Pyrrolidines as Potent Functional Agonists of the Human Melanocortin-4 Receptor. *Bioorg. Med. Chem. Lett.* **2007,** *17*, 5165–5170.

5. Tran, J. A.; Tucci, F. C.; Jiang, W.; Marinkovic, D.; Chen, C. W.; Arellano, M.; Markison, S.; Fleck, B. A.; Wen, J.; White, N. S.; Pontillo, J.; Saunders, J.; Marks, D.; Hoare, S. R.; Madan, A.; Foster, A. C.; Chen, C.; Tran, J. A.; Tucci, F. C.; Jiang, W.; Marinkovic, D.; Chen, C. W.; Arellano, M.; Markison, S.; Fleck, B. A.; Wen, J.; White, N. S.; Pontillo, J.; Saunders, J.; Marks, D.; Hoare, S. R.; Madan, A.; Foster, A. C.; Chen, C., Pyrrolidinones as Orally Bioavailable Antagonists of the Human Melanocortin-4 Receptor with Anti-Cachectic Activity. *Bioorg. Med. Chem. Lett.* **2007,** *15*, 5166–5176.
6. Maggiora, G. M.; Maggiora, G. M., On Outliers and Activity Cliffs–Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006,** *46*, 1535–1535.
7. Guha, R.; Van Drie, J. H., The Structure-Activity Landscape Index: Identifying and Quantifying Activity-Cliffs. *J. Chem. Inf. Model.* **2008,** *48* (3), 646–658.
8. Bertz, S., The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981,** *103* (12), 3599–3601.
9. Allu, T. K.; Oprea, T. I., Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *J. Chem. Inf. Model.* **2005,** *45* (5), 1237–1243.
10. Peltason, L.; Bajorath, J., SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007,** *50*, 5571–5578.
11. Lehninger, A. L., *Lehninger Principles of Biochemistry.* 4th ed.; W.H. Freeman: New York, 2004.
12. Guha, R.; Van Drie, J. H., Assessing How Well a Modeling Protocol Captures a Structure-Activity Landscape. *J. Chem. Inf. Model.* **2008,** *48* (8), 1716–1728.
13. Duch, W., *Feature Extraction: Foundations and Applications.* Springer: Berlin, 2006; Vol. 207.
14. Takahashi, H.; Bekkali, Y.; Capolino, A. J.; Gilmore, T.; Goldrick, S. E.; Kaplita, P. V.; Liu, L.; Nelson, R. M.; Terenzio, D.; Wang, J.; Zuvela-Jelaska, L.; Proudfoot, J.; Nabozny, G.; Thomson, D.; Takahashi, H.; Bekkali, Y.; Capolino, A. J.; Gilmore, T.; Goldrick, S. E.; Kaplita, P. V.; Liu, L.; Nelson, R. M.; Terenzio, D.; Wang, J.; Zuvela-Jelaska, L.; Proudfoot, J.; Nabozny, G.; Thomson, D., Discovery and SAR Study of Novel Dihydroquinoline Containing Glucocorticoid Receptor Agonists. *Bioorg. Med. Chem. Lett.* **2007,** *17*, 5091–5095.
15. Takahashi, H.; Bekkali, Y.; Capolino, A. J.; Gilmore, T.; Goldrick, S. E.; Nelson, R. M.; Terenzio, D.; Wang, J.; Zuvela-Jelaska, L.; Proudfoot, J.; Nabozny, G.; Thomson, D.; Takahashi, H.; Bekkali, Y.; Capolino, A. J.; Gilmore, T.; Goldrick, S. E.; Nelson, R. M.; Terenzio, D.; Wang, J.; Zuvela-Jelaska, L.; Proudfoot, J.; Nabozny, G.; Thomson, D., Discovery and SAR Study of Novel Dihydroquinoline Containing Glucocorticoid Receptor Ligands. *Bioorg. Med. Chem. Lett.* **2006,** *16*, 1549–1552.
16. Erdos, P.; Renyi, A., On Random Graphs. *Publ Math.* **1959,** *6*, 290–297.
17. Freeman, L. C., Centrality in Social Networks I: Conceptual Clarification. *Social Networks* **1979,** *1*, 215–239.
18. Wasserman, S.; Faust, K., *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press, 1994.
19. Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J., Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008,** *51* (19), 6075–6084.
20. Wawer, M.; Peltason, L.; Bajorath, J., Elucidation of Structure-Activity Relationship Pathways in Biological Screening Data. *J. Chem. Inf. Model.* **2009,** *52* (4), 1075–1080.

Chapter 4

Computational Analysis of Activity and Selectivity Cliffs

Lisa Peltason and Jürgen Bajorath

Abstract

The exploration of structure–activity relationships (SARs) is a major challenge in medicinal chemistry and usually focuses on compound potency for individual targets. However, selectivity of small molecules that are active against related targets is another critical parameter in chemical lead optimization. Here, an integrative approach for the systematic analysis of SARs and structure–selectivity relationships (SSRs) of small molecules is presented. The computational methodology is described and a cathepsin inhibitor set is used to discuss key aspects of the analysis. Combining a numerical scoring scheme and graphical visualization of molecular networks, the approach enables the identification of different local SAR and SSR environments. Comparative analysis of these environments reveals variable relationships between molecular structure, potency, and selectivity. Furthermore, key compounds are identified that are involved in the formation of activity and/or selectivity cliffs and often display structural features that determine compound selectivity.

Key words: Active compounds, Target selectivity, Structure–activity relationships, Structure–selectivity relationships, Activity cliff, Selectivity cliff

1. Introduction

It is a central paradigm in medicinal chemistry that similar molecules should also share similar biological activity. This intuitive "similarity–property principle" [1] has been substantiated by a wealth of observations and continues to be widely accepted in the medicinal chemistry community. However, hit-to-lead or lead optimization stages of a drug discovery program often benefit from the contrary situation: active compounds are subjected to minor chemical modifications to significantly improve potency and/or selectivity. Such optimization efforts are most effective in the presence of "activity cliffs" that are marked by highly similar compounds having dramatic differences in potency [2]. Activity cliffs are thought to result from the presence or absence of

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_4, © Springer Science+Business Media, LLC 2011

structural patterns that are required for biological activity and are indicative of "discontinuous" structure–activity relationships (SARs). This SAR phenotype is characterized by large-magnitude biological responses to chemical modifications of active molecules. By contrast, SARs are considered "continuous" if chemical alterations result in only gradual potency changes. Consistent with the similarity–property principle, this SAR type is also frequently encountered in medicinal chemistry. It has been shown that these two elementary SAR categories are not mutually exclusive and often coexist within compound activity classes [3], giving rise to "heterogeneous" SARs. Furthermore, the nature of SARs can be highly complex, depending on the types of molecules under study [3]. Hence, it is often not clear whether a given compound class might be susceptible to chemical optimization. In view of these observations, computational approaches to the systematic analysis of SARs gain increasing importance [4]. Methods to study SARs of compound activity classes on a large scale, determine global and local SAR features, and identify or quantify activity cliffs have recently been introduced [5].

While these methods typically focus on target-specific compound potency, this is only one of several critical factors that need to be considered in lead optimization. For example, besides high potency, a promising drug candidate must also have a desired selectivity profile against a number of targets and antitargets. Target selectivity of active compounds is not always the result of exclusive target binding events but often emerges from differential potency profiles against multiple targets [6]. The resulting multitarget SARs ultimately give rise to structure–selectivity relationships (SSR), which are particularly relevant for target families of closely related proteins. Similar to SARs, the nature of SSRs is essentially determined by the way active compounds respond to chemical alterations. In a recent study, it has been demonstrated that SSRs can be classified into fundamentally different phenotypes in analogy to SARs [7]. Continuous SSRs are characterized by the presence of active compounds of increasing structural diversity that share similar selectivity for related targets. By contrast, the major characteristic of discontinuous SSRs is the presence of "selectivity cliffs" formed by similar molecules with markedly different selectivity behavior.

Although there has been increasing interest in computational approaches to analyze and predict selectivity, the computational study of SSRs is still in its infancy [8]. In this chapter, we present a computational methodology aimed at the systematic analysis of SAR and SSR features and the detection of activity and selectivity cliffs. Taking into account that SSRs are often the consequence of variable SARs against multiple targets, an integrated approach is applied to study single-target SARs and SSRs for pairs of related targets. A scoring scheme termed SAR Index (SARI; [9]) is

utilized to classify SAR and SSR features on the basis of compound activity classes, series of related compounds, or individual molecules. The detection of activity and selectivity cliffs is facilitated through graphical representation in network-like similarity graphs (NSGs; [10]). These graphs display similarity relationships and potency or selectivity distributions within an activity class and make it possible to identify regions of different local SAR or SSR character. Focusing on discontinuous local SAR or SSR environments enables the exploration of activity or selectivity cliffs and molecular determinants of compound potency and selectivity.

2. Materials and Methods

In the following, we describe an integrated approach for the detailed analysis of single-target SAR and target-pair SSR features at different levels of detail. On the basis of a numerical scoring function, SARs and SSRs are quantitatively assessed at the level of compound activity classes, compound subsets identified through similarity-based clustering, and individual molecules. A graphical representation technique complements the methodology.

2.1. Potency and Selectivity Data

For the comparative analysis of SAR and SSR features in a set of active molecules, biological potency measurements for two related targets are required. Potency data, usually available either as IC_{50} or K_i values, should be converted to pIC_{50} or pK_i values by calculating the negative logarithm to the base of 10 of the original potency measurements. Selectivity values for target A over another target B are then determined as the difference between the corresponding logarithmic potency values of each compound:

$$S_i = P_i(A) - P_i(B).$$

Here, S_i stands for the selectivity value of compound i for target A over target B and $P_i(A)$ and $P_i(B)$ denote its potency values for targets A and B, respectively.

2.2. Molecular Similarity Calculation

The analysis of SARs or SSRs requires a measure to systematically evaluate chemical modifications through pairwise comparison of molecular structures. The methodology presented herein makes use of two-dimensional molecular similarity calculated on the basis of MACCS structural keys [11]. The binary MACCS key fingerprint consists of 166 bits that monitor the presence or absence of 166 predefined structural features. If a specific substructure is found in a molecule, the corresponding bit is set to 1 ("on"); otherwise, it is set to 0 ("off"). The similarity between two molecules is then determined by comparison of their

fingerprint representations. Here, the Tanimoto coefficient (*Tc*) is utilized to calculate MACCS fingerprint similarity. The *Tc* presents a measure of bit string overlap and is defined as follows for two binary fingerprints *i* and *j*:

$$Tc(i,j) = \frac{N_{ij}}{N_i + N_j - N_{ij}}.$$

Here, N_{ij} is the number of bits that are set on in both fingerprints and N_i and N_j refer to the number of bits that are set on in *i* and *j*, respectively. Given this formulation, identical fingerprints obtain a maximal *Tc* value of 1, whereas nonoverlapping fingerprints are assigned a *Tc* value of 0. Fingerprint representations were calculated using the Molecular Operating Environment (MOE; [12]).

2.3. Assessment of Global SAR and SSR Discontinuity

For the quantitative assessment of continuous and discontinuous SAR elements, a scoring function termed SAR Index (SARI) has been introduced [9]. SARI is calculated on the basis of similarity and potency values within a given activity class and combines two individual score components. While the "continuity score" takes intraclass structural diversity and compound potency into account as an indicator of continuous SARs, the "discontinuity" score estimates the presence of activity cliffs within a data set. The methodology presented herein aims to analyze SAR and SSR discontinuity and associated activity and selectivity cliffs; hence, we will focus on the discontinuity score. The original, potency-based discontinuity score is defined as the average pairwise potency difference between pairs of similar compounds in an activity class *A*, multiplied with pairwise similarity:

$$\mathrm{disc}_{\mathrm{raw}}(A) = \underset{\left\{\begin{smallmatrix}(i,j)\in A|\mathrm{sim}(i,j)>0.65,\\ |P_i-P_j|>1\end{smallmatrix}\right\}}{\mathrm{mean}} (|P_i - P_j| \cdot \mathrm{sim}(i,j)).$$

Here, only compound pairs are considered that exceed a predefined similarity threshold, which is typically set to 0.65 for MACCS keys but can be adjusted according to data set size and composition. Furthermore, for a compound pair to be considered in discontinuity score calculations, we require a potency difference of more than one order of magnitude in order to focus on significant activity cliffs. Multiplication of the potency difference with pairwise compound similarity puts high emphasis on potency differences between highly similar molecules.

The discontinuity score has also been adapted for the assessment of SSR discontinuity within a compound set. For this purpose, selectivity values calculated as described above are utilized instead of potency values. This formulation is appropriate because selectivity values are derived from potency differences and can thus be utilized in the same manner.

The calculation of potency- and selectivity-based discontinuity scores for entire compound activity classes monitors the presence of activity and selectivity cliffs within a data set and provides a measure of global SAR and SSR discontinuity. High discontinuity score values indicate a high degree of SAR/SSR discontinuity and the presence of significant potency or selectivity differences between structurally similar molecules.

2.4. Local SAR and SSR Assessment

In order to identify series of compounds within a data set that are characterized by distinct SARs or SSRs, activity classes are divided into subsets of similar molecules by means of similarity-based clustering. The molecules of an activity class are subjected to hierarchical clustering using their pairwise MACCS *Tc* similarity values and Ward's minimum variance linkage method [13]. For each of the resulting compound clusters, potency- and selectivity-based discontinuity scores are calculated as described above. Thus, local SAR and SSR features in compound subsets are quantified, with high cluster discontinuity scores indicating subsets of similar compounds that have markedly different potency or selectivity levels.

2.5. Identification of Activity and Selectivity Cliffs

For the detailed analysis of activity and selectivity cliffs, a modified version of the discontinuity score has been developed that is calculated on the basis of individual compounds. The aim is to focus on compounds that are involved in the formation of activity and/or selectivity cliffs. To this end, the compound discontinuity score accounts for potency and selectivity differences between a given active molecule and all molecules that are similar to it (again applying a MACCS *Tc* similarity threshold of 0.65). In contrast to global score calculations, no potency difference cutoff is required here because for the assessment of discontinuity contributions from individual compounds, all "neighbors" (i.e., similar compounds) of a molecule must be taken into account. Hence, for a given molecule i in an activity class A, the potency-based compound discontinuity score is defined as

$$\mathrm{disc}_{\mathrm{raw}}(i) = \underset{\{j \in A | j \neq i, \mathrm{sim}(i,j) > 0.65\}}{\mathrm{mean}} (|P_i - P_j| \cdot \mathrm{sim}(i,j)).$$

The selectivity-based compound score is defined analogously. This function assigns high scores to molecules that have potency or selectivity significantly different from their structural neighbors and are thus involved in the formation of activity or selectivity cliffs.

2.6. Score Normalization

For ease of comparison, the discontinuity scores for entire data sets, compound clusters, and individual compounds are standardized and normalized to the value range between 0 and 1. For normalization of global activity class and cluster scores, a panel of

13 reference activity classes assembled from the MDDR [14] is taken as a basis and the "raw" scores of each class are normalized with respect to the score distribution within this reference panel, as described in the following.

Initially, the sample mean ($\overline{\text{disc}_{\text{raw}}}$) and sample standard deviation (s_{disc}) of the scores within the set of reference classes are calculated. These reference values are then used to calculate Z-scores from the raw scores of each activity class or compound cluster A:

$$\text{disc}_{\text{zscore}}(A) = \frac{\text{disc}_{\text{raw}} - \overline{\text{disc}_{\text{raw}}}}{s_{\text{disc}}}.$$

Finally, the standardized scores are mapped onto the value range [0,1] by calculating the value of the cumulative distribution function for each Z-score under the assumption of a normal distribution:

$$\text{disc}_{\text{norm}}(A) = \Theta(\text{disc}_{\text{zscore}}(A)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{disc}_{\text{zscore}}(A)} \exp(-\frac{1}{2}x^2)\text{d}x.$$

This function indicates for a given Z-score value the probability of the event that the Z-score of a randomly chosen class is less than or equal to this value. Hence, a Z-score of 0 obtains a normalized value of 0.5 because it corresponds to the mean of the entire raw score distribution. Increasing Z-score values obtain values closer to 1, and decreasing Z-scores approach a value of 0. Hence, normalized discontinuity score values near 0 correspond to a low degree of SAR or SSR discontinuity, whereas values near 1 indicate the opposite situation. Applying a common normalization scheme for potency- and selectivity-based scores permits the direct comparison of SAR and SSR features at the level of compound data sets and clusters and makes it possible to relate SSR features to the previously established SAR categories.

Discontinuity scores on the basis of individual compounds are also normalized to adopt values between 0 and 1 by calculation of Z-scores and the cumulative distribution function. However, compound discontinuity scores are standardized relative to all compound scores within the activity class under study. Thus, no external reference set is required. This normalization scheme scales the compound scores to the score distribution within a given activity class, which makes it possible to identify key compounds that make the largest discontinuity contributions in the activity class. It should be noted, however, that this normalization procedure does not permit the comparison of compound scores across different classes. Furthermore, potency-based and selectivity-based compound scores are normalized separately; i.e., potency-based scores are normalized with respect to the

potency-based compound score distribution in a data set, and selectivity-based scores take the distribution of selectivity-based compound scores as a reference.

2.7. Graphical Representation

The scoring framework described above provides multilayered information concerning the SAR or SSR character of compound activity classes. NSGs have been designed to visualize these different levels of information and make it possible to relate different SAR and SSR features to each other in an intuitive manner. In these graphs, compounds of an activity class are represented as nodes and edges between them account for similarity relationships. For an activity class annotated with biological data for two targets *A* and *B*, three different NSG representations can be generated: two potency-based graphs NSG_A and NSG_B that utilize potency values for the two targets, respectively, and a selectivity-based graph NSG_{AB} utilizing calculated selectivity values for target *A* over target *B*. Figure 1 schematically illustrates the individual components of an NSG. Five different levels of information can be distinguished. Firstly, similarity relationships between molecules are reflected by edges that connect two nodes if the corresponding molecules exceed a predefined similarity threshold value (here MACCS *Tc* of 0.65). Secondly, the potency or selectivity distribution within an activity class is represented by node shading. Nodes are gray-shaded according to potency or selectivity values of the corresponding compounds using a gradient from black to white, with black indicating highest, and white lowest values within a class. In potency-based NSGs, a common grayscale is applied for both targets, ranging from the lowest to the highest potency of a compound active against one or the other target. In selectivity-based NSGs, the spectrum ranges from the highest observed selectivity for one target (black) to the corresponding inverse selectivity value for the other target (white). Accordingly,

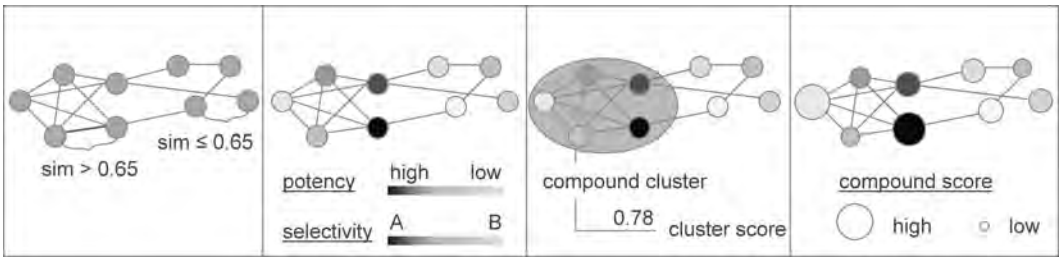


Fig. 1. Schematic representation of a Network-like Similarity Graph (NSG). Compounds are represented as nodes that are connected by edges if their pairwise similarity exceeds a predefined threshold value. Nodes are *gray-shaded* according to potency or selectivity. Sets of similar compounds are identified through similarity-based clustering and characterized by local discontinuity scores. Furthermore, discontinuity scores are calculated on a per-compound basis and nodes are scaled in size according to the magnitude of the scores. High compound scores indicate key compounds that form activity and/or selectivity cliffs.

nonselective compounds that have similar potency for both targets are represented by intermediate shades of gray. A third level of information is presented by compound clusters that indicate subsets of similar molecules. The fourth level of information is provided by discontinuity scores of compound clusters, which highlight local environments of different SAR and SSR character within a compound set. Finally, at the fifth level of information, compound discontinuity scores reveal contributions to overall SAR and SSR discontinuity made by individual compounds. Therefore, nodes are scaled in size according to compound discontinuity scores, with the largest nodes corresponding to compounds that are involved in the formation of the most significant activity and/or selectivity cliffs.

3. Application Example

In the following, key aspects of the methodology are discussed on the basis of an exemplary selectivity data set. A total of 159 inhibitors of cathepsin (cat) L and B were taken from previously published compound sets [15]. Figure 2 shows the NSG representations for the inhibitor set based on potency for cat L (Fig. 2a) and cat B (Fig. 2b) as well as selectivity values for cat L over cat B, calculated from the potency difference as described above (Fig. 2c). It should be noted that the topology of the networks is determined on the basis of compound similarity relationships and is thus the same for all three networks.

Potency- and selectivity-based discontinuity scores for this compound set are overall low to intermediate with values of 0.41 for cat L, 0.26 for cat B, and 0.33 for cat L over cat B. However, individual compound clusters can be identified that display a distinct degree of local discontinuity, as indicated by high cluster discontinuity scores. Cluster score values within the networks cover essentially the entire range from 0 to 1, which indicates the coexistence of different local SAR or SSR environments and is a hallmark of heterogeneous SARs or SSRs. Moreover, individual clusters might obtain substantially different scores in the potency- and selectivity-based NSGs, representing environments of different local single-target SAR and dual-target SSR character, as in the case of the clusters labeled “A,” “B,” and “C” in Fig. 2. In all three NSG representations, discontinuous network regions contain compounds that obtain high compound discontinuity scores, represented as large nodes in the networks. These compounds are involved in the formation of activity or selectivity cliffs and introduce SAR or SSR discontinuity into the activity landscape. Accordingly, they can be considered key compounds that largely determine SAR and SSR features within an

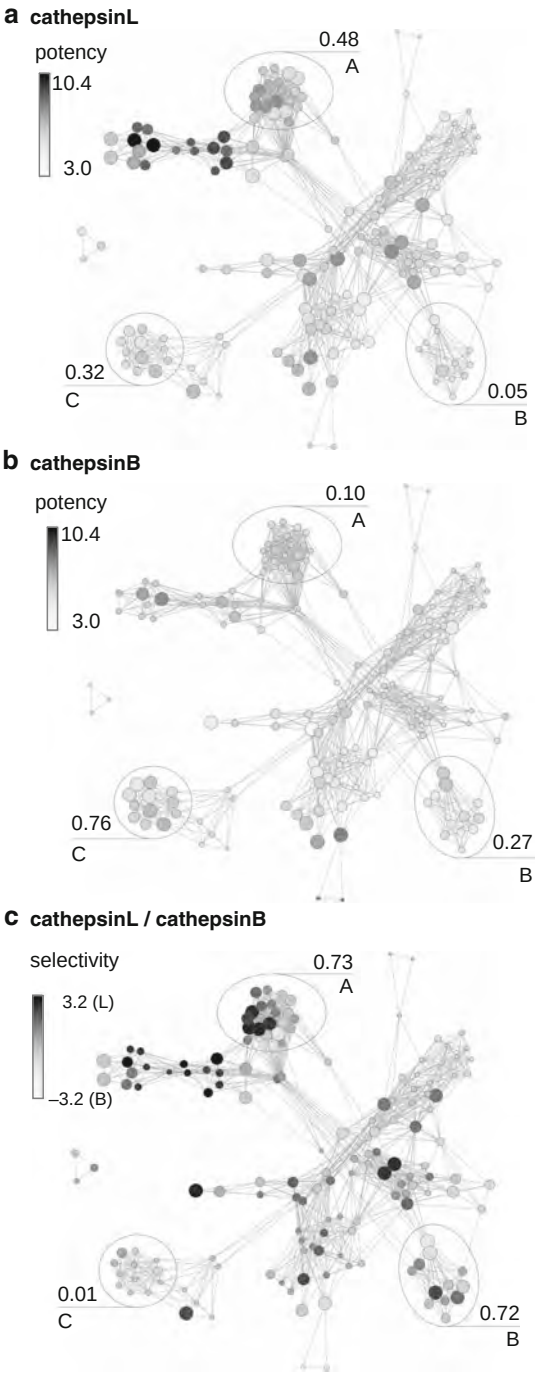


Fig. 2. NSG representations for 159 inhibitors of cathepsin (cat) L and B. Labeled compound clusters are discussed in the text. (**a**) NSG based on potency values for cat L. (**b**) NSG based on potency values for cat B. (**c**) NSG based on selectivity values for cat L over cat B.

activity class. In Fig. 2, key compounds are found in various regions of the networks but are not distributed equally in the potency and selectivity NSGs. Hence, a compound might represent a key compound in one NSG but not in another. Focusing on such molecules that contribute to SAR and SSR discontinuity in different ways helps to better understand relationships between molecular structure, potency, and selectivity, and ultimately leads to the identification of molecular selectivity determinants.

For example, cluster "A" presents an environment of different local SAR and SSR character, with low to moderate potency-based discontinuity scores for cat L (0.48) and cat B (0.10) and a high degree of selectivity-based discontinuity for cat L over B (0.73). These different levels of discontinuity can be explained on the basis of differences in potency and selectivity distributions within this cluster. As indicated by node shading in Fig. 2a and b, potency for cat B is distributed more homogeneously than for cat L, which results in overall lower potency differences among similar compounds and is reflected by a lower discontinuity score. Furthermore, compounds within this cluster have very different levels of selectivity against the two targets including L- and B-selective molecules, as indicated by the variable node shading in Fig. 2c. Several molecules can be identified that obtain high compound discontinuity scores and are associated with activity and selectivity cliffs. Figure 3 shows a pair of molecules that are structural analogs and form activity and selectivity cliffs of increasing magnitude. For cat B, the two analogs have a potency

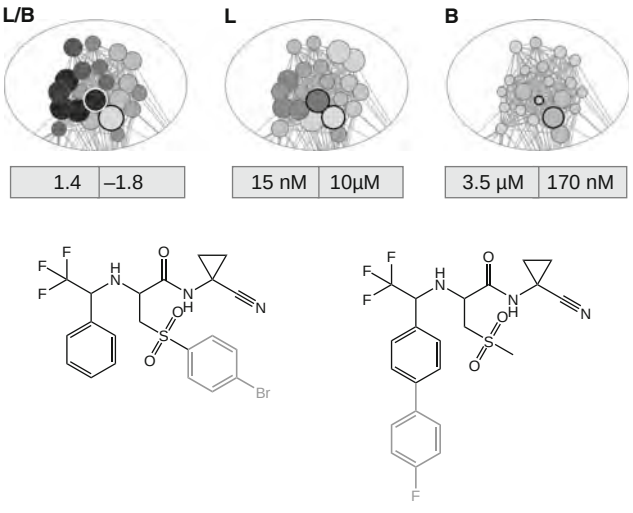


Fig. 3. Activity and selectivity cliff markers. Shown is a pair of structural analogs that have significantly different potency and selectivity against cat L and cat B, thus forming activity and selectivity cliffs. Network details present cluster "A" from Fig. 4.2. Nodes encircled in *black* or *white* indicate the shown inhibitors.

difference of more than one order of magnitude and hence form a moderate activity cliff. The potency difference is even larger for cat L, which gives rise to a significant activity cliff for this target. The two compounds also display significantly different selectivity, one compound being selective for cat L and the other for cat B, which results in a selectivity cliff. In this case, the location of a halogenated phenyl substituent determines whether a compound is selective for cat L or for cat B. Hence, this pair of compounds represents activity cliff markers that also form a selectivity cliff and strongly determine local SAR and SSR discontinuity.

Cluster "B" also presents a network region that displays different local SAR and SSR character. In this case, however, no significant degree of SAR discontinuity is observed for either target, consistent with low potency-based cluster score values of 0.05 for cat L and 0.27 for cat B, respectively. By contrast, the degree of discontinuity in target-pair SSRs is much higher than in single-target SARs, as indicated by a selectivity-based discontinuity score of 0.72. Consistent with this observation, the potency distribution within this cluster is more homogeneous than the distribution of selectivity values, and more key compounds (represented as large nodes) are found in this region of the selectivity-based NSG than in the corresponding potency-based network regions. Figure 4 shows an exemplary pair of analogs from this cluster that are distinguished only by a chlorine substituent at the benzene ring and have markedly different selectivity values. The chlorine-substituted analog is selective for cat L, whereas the nonsubstituted compound is nonselective with a tendency toward cat B. However,

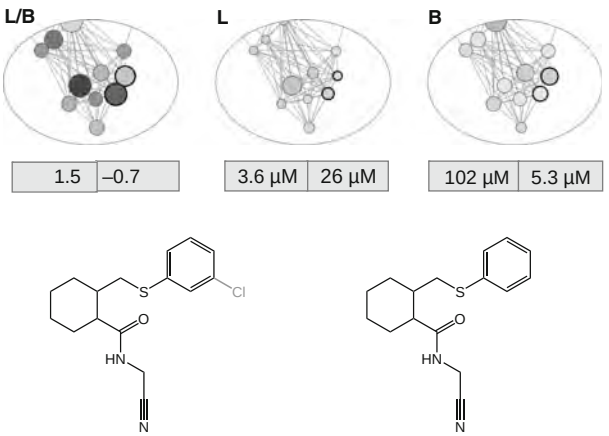


Fig. 4. Selectivity cliff markers. Another pair of cat L and cat B inhibitors is shown that consists of structural analogs that have significantly different selectivity levels and hence form a selectivity cliff. The potency differences are less distinct in this case. Network details correspond to cluster "B" from Fig. 4.2. *Encircled nodes* indicate the displayed inhibitors.

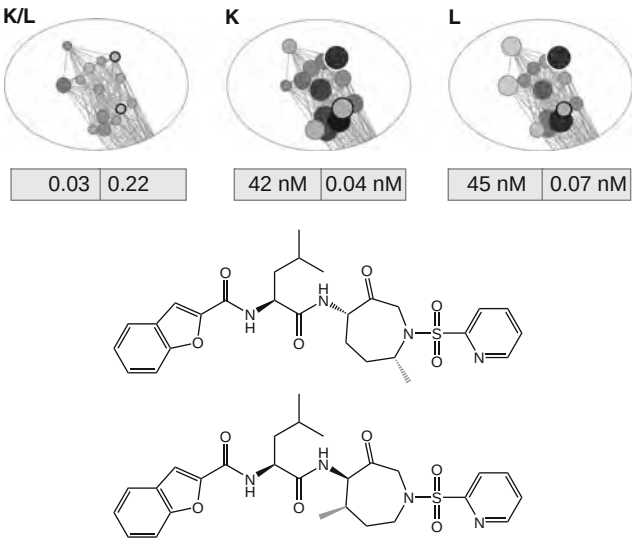


Fig. 5. Activity cliff markers. Two analogs from another cathepsin inhibitor set are shown that inhibit cat K and cat L at significantly different levels, but are both non-selective with respect to the two targets. Therefore, these compounds form activity cliffs, but no selectivity cliff. Network details present their environment within the NSG representations of the cat K/L data set (not shown). *Encircled nodes* indicate the shown inhibitors.

potency values of the two compounds are comparable for individual targets. Thus, these compounds present selectivity cliff markers that are not associated with significant activity cliffs and contribute to SAR and SSR discontinuity in different ways.

However, compounds that form activity cliffs are not always involved in the formation of selectivity cliffs. Figure 5 presents a pair of cathepsin K and L inhibitors from another inhibitor set [15]. These molecules belong to a compound cluster that displays remarkable SAR discontinuity but essentially no SSR discontinuity, reflected by cluster discontinuity scores of 0.86 and 0.81 using potency against cat K and cat L, respectively, and 0.05 for selectivity for cat K over cat L. This different SAR and SSR behavior results from the presence of molecules that have distinct potency differences for individual targets, but respond to both targets in a similar way, as illustrated in Fig. 5. The molecules are distinguished only by the position of a methyl group at the central azepane ring but have markedly different potency against cat K and L. Thus, they form significant activity cliffs for both targets. However, each of the analogs has similar potency levels for cat K and cat L, which renders them nonselective. Accordingly, the position of the methyl group that distinguishes these compounds can be considered important for compound potency but not for selectivity.

4. Conclusions

The analysis of target selectivity of compounds that are active against multiple targets is of critical importance for lead optimization and drug discovery. Simultaneous optimization of compound potency and selectivity is a major challenge for medicinal chemistry and requires a thorough analysis of structure–activity and structure-selectivity relationships. For this purpose, an integrated multicomponent methodology has been developed that combines a numerical scoring scheme with graphical network representations. Comparative analysis of potency- and selectivity-based molecular networks makes it possible to identify series of compounds that are characterized by distinct local SARs and SSRs. These local SAR and SSR features might differ substantially, which illustrates the complementary nature of potency and selectivity information and the intrinsic variability of SARs and SSRs. Focusing on regions of local discontinuity leads to the identification of molecules that are involved in the formation of activity and/or selectivity cliffs and significantly influence SAR and SSR features. Systematic exploration of these key compounds and their local network environments often reveal structural patterns that determine compound potency and selectivity.

The approach has considerable practical utility for medicinal chemistry, because it permits the selection of compound series based on SAR and/or SSR behavior. Discontinuous local SARs and SSRs are considered particularly promising for chemical optimization. In these cases, potency and selectivity are likely to change (and, hopefully, improve) significantly in response to chemical modifications. Furthermore, the analysis of activity and selectivity cliffs often leads to a better understanding of relationships between molecular structure, potency, and selectivity, which is a prerequisite for successful compound optimization.

References

1. Johnson, M. A. and Maggiora, G. M., eds. (1990) Concepts and applications of molecular similarity. New York, NY: Wiley.
2. Maggiora, G. M. (2006) On outliers and activity cliffs – Why QSAR often disappoints *J Chem Inf Model* **46**, 1535.
3. Peltason, L. and Bajorath, J. (2007) Molecular similarity analysis uncovers heterogeneous structure–activity relationships and variable activity landscapes *Chem Biol* **14**, 489–497.
4. Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., and van Drie, J. (2009) Navigating structure–activity landscapes *Drug Discov Today* **14**, 698–705.
5. Peltason, L. and Bajorath, J. (2009) Systematic computational analysis of structure–activity relationships: Concepts, challenges, and recent advances *Future Med Chem* **1**, 451–466.
6. Hopkins, A. L. (2008) Network pharmacology: The next paradigm in drug discovery *Nat Chem Biol* **4**, 682–690.
7. Peltason, L., Hu, Y., and Bajorath, J. (2009) From structure–activity to structure–selectivity relationships: Quantitative assessment,

selectivity cliffs, and key compounds *Chem Med Chem* **4**, 1864–1873
8. Bajorath, J. (2008) Computational analysis of ligand relationships within target families *Curr Opin Chem Biol* **12**, 352–358.
9. Peltason, L. and Bajorath, J. (2007) SAR index: Quantifying the nature of structure–activity relationships *J Med Chem* **50**, 5571–5578.
10. Wawer, M., Peltason, L., Weskamp, N., Teckentrup, A., and Bajorath, J. (2008) Structure–activity relationship anatomy by network-like similarity graphs and local structure–activity relationship indices *J Med Chem* **51**, 6075–6084.
11. MACCS Structural Keys: Symyx Software, San Ramon, CA (USA).
12. MOE (Molecular Operating Environment): Chemical Computing Group Inc., Montreal, Quebec (Canada).
13. Ward, J. H. (1963) Hierarchical grouping to optimize an objective function *J Am Stat Assoc* **58**, 236–244.
14. MDDR (MDL Drug Data Report): Symyx Software, San Ramon, CA (USA).
15. Stumpfe, D., Geppert, H., and Bajorath, J. (2008) Methods for computer-aided chemical biology. Part 3: Analysis of structure–selectivity relationships through single- or dual-step selectivity searching and Bayesian classification *Chem Biol Drug Des* **71**, 518–528.

Chapter 5

Similarity Searching Using 2D Structural Fingerprints

Peter Willett

Abstract

This chapter reviews the use of molecular fingerprints for chemical similarity searching. The fingerprints encode the presence of 2D substructural fragments in a molecule, and the similarity between a pair of molecules is a function of the number of fragments that they have in common. Although this provides a very simple way of estimating the degree of structural similarity between two molecules, it has been found to provide an effective and an efficient tool for searching large chemical databases. The review describes the historical development of similarity searching since it was first described in the mid-1980s, reviews the many different coefficients, representations, and weightings that can be combined to form a similarity measure, describes quantitative measures of the effectiveness of similarity searching, and concludes by looking at current developments based on the use of data fusion and machine learning techniques.

Key words: Chemical databases, Chemoinformatics, Data fusion, Fingerprint, Fragment substructure, Machine learning, Similar property principle, Similarity coefficient, Similarity measure, Similarity searching, Weighting scheme

1. Introduction

The Collins English Dictionary defines similar to be "showing resemblance in qualities, characteristics or appearance; alike but not identical" and the comparison of objects to determine their levels of similarity lies at the heart of many academic disciplines. Thus, archaeologists may study the relationships between pot shards from different historical sites; literary studies may involve comparing fragments of poetry from different works by – possibly – the same author; and modern systematics derives from the attempts of the medieval apothecaries to group medicinal plants. The definitions of similarity, and the purposes for which these definitions are employed, in these three applications are very different, but they have in common the aim of synthesising new knowledge from a similarity-based analysis of that which already exists. Similarity

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_5, © Springer Science+Business Media, LLC 2011

concepts have long played an important role in chemistry [1]; indeed one of the most striking examples is the work of Mendeleev that led to the establishment of the modern Periodic Table, by means of which he was able not only to classify the existing elements but also to predict the existence of elements that were then unknown.

In this chapter, we focus on one specific application of similarity in chemoinformatics: *similarity searching*, i.e., the ability to scan through a database of chemical molecules to find those that are most similar to some user-defined query molecule [2–7]. In what follows, we shall normally refer to the query as the *reference structure*; an alternative name that is frequently used in the literature is the *target structure*, but we believe that the former name is to be preferred given the possibility of confusion with a biological target.

Similarity searching is one particular type of *virtual screening*. This is the use of a computational technique for selecting molecules for subsequent investigation, most obviously for testing for bioactivity in a lead-discovery programme. There are many different virtual screening methods available, but they all have the common aim of ranking a list of possible molecules so that those at the top of the ranking are expected to have the greatest probability of activity. Virtual screening methods differ in the amount of information that is available [8–13]. Similarity searching has by far the smallest information requirement, needing just a single known bioactive molecule (as discussed further below). Examples of other approaches to virtual screening include: 2D or 3D substructure searching (which require the availability of a topological or geometric pharmacophore, respectively, these being derived from a small number of known bioactive molecules); machine learning methods (which require large numbers of both known active and known inactive molecules); and docking methods (which require the 3D structure of the biological target). The many methods that are now available have led to comparisons that seek to determine the relative effectiveness of different approaches to screening; the reader is referred to the literature for discussions of the strengths and weaknesses of similarity searching as compared to other screening approaches (*see*, e.g., [7, 14–20]).

This chapter seeks to present the basic principles of similarity searching, eschewing detailed discussion of individual approaches, and is structured as follows. Subsection 2 provides an introduction to similarity searching, and describes the *similar property principle* that underlies the use of similarity as a tool for database searching. Subsection 3 discusses the three components – the *representation*, the *similarity coefficient*, and the *weighting scheme* – that comprise a *similarity measure* for computing the degree of resemblance between two molecules; the focus of this chapter is one particular type of representation, the 2D *fingerprint*, and this representation

is hence discussed in some detail in this section. Subsection 4 discusses the criteria that have been used to evaluate the retrieval effectiveness of different types of similarity searching procedure. Finally, Subsection 5 summarises recent work that involves the use of not just a single reference structure, as is used for conventional similarity searching, but multiple reference structures.

The coverage of this review is intentionally focused, considering only one representation of molecular structure (the 2D *fingerprint*) and only one application of similarity (similarity searching). The reader is referred to the literature for more general discussions of chemoinformatics [21–23] and of other similarity-related topics, such as 3D similarity measures, cluster analysis, molecular diversity analysis, and reaction similarity [3, 24–26]; for additional material specifically about similarity searching, it is worth noting that a characteristic of the field is that much of the work has been carried out by a limited number of research groups, most notably those directed by Bajorath [27], Sheridan [16], and Willett [28].

2. The Similar Property Principle

The input to a similarity search is the reference structure for which related structures are required. In the drug-discovery context, the reference structure normally exhibits a potentially useful level of biological activity and might be, for example, a competitor's compound or a structurally novel hit from an initial high-throughput screening (HTS) experiment. Thus, the reference structure is normally an entire molecule, rather than the partial structure that forms the basis for 2D or 3D substructure searching (that said, there has been some interest in similarity searches of molecules that are substructures or superstructures of the reference structures [29–31]). Each database structure is encoded using the same representation scheme as was used to encode the reference structure; the two representations are compared to ascertain the level of structural commonality using a similarity coefficient. In some cases, a weighting scheme is applied to one or both of the representations prior to the calculation of the similarity, with the aim of increasing the relative importance of particular features within the overall representation. The similarities are computed in this way for every molecule in the database that is being searched, and then the similarity values sorted into descending order. The molecules at the top of the resulting ranking, which are often referred to as the *nearest neighbours* as they are the closest in some sense to the reference structure, are then presented to the user as the output from the similarity search.

This approach to database access was first described by Carhart et al. [32] and by Willett et al. [33]. Both of these studies found

that effective measures of chemical similarity could be obtained by determining the numbers of 2D substructures common to a reference structure and a database structure, although the starting points for the two studies were rather different. Carhart et al., working at Lederle Laboratories, used the information about common fragments not just for similarity searching but also for substructural analysis (vide infra). The study by Willett et al. drew on earlier work by Adamson and Bush that reported probably the very first use of 2D fingerprints for the calculation of molecular similarity (specifically in the context of QSAR studies rather than for large-scale database applications) [34]. Willett et al. used the information about common fragments in a combined search system at Pfizer, where the computed similarities were used to rank the molecules retrieved in a substructure search; however, the authors soon realised that the initial substructure search was not necessary and that similarity searching on its own provided a novel way of accessing a chemical database.

Following these two initial studies, fragment-based similarity searching was adopted very rapidly in both commercial and in-house chemoinformatics systems. Its uptake was spurred by several factors: it provides a retrieval mechanism that is complementary to substructure searching; it uses the same basic data as existing substructure software, i.e., sets of 2D fingerprints; and it is both rapid and powerful in execution, encouraging interactive exploration of the range of structural types in a database [35]. These are all perfectly valid, but essentially pragmatic reasons for using similarity searching. There is, however, also a rational basis, which derives from what is known as the Similar Property Principle. The Principle states that molecules that have similar structures will have similar properties, and is normally ascribed to Johnson and Maggiora, whose 1990 book was the first to highlight the role of similarity in what we now refer to as chemoinformatics [25]. However, it had certainly been discussed prior to then, e.g., by Wilkins and Randic in 1980 [36], and arguably underlies the whole area of drug discovery: if there was not some relationship between molecular structures (however, these are represented in computational terms) and molecular properties, then lead discovery and lead optimisation would be essentially random processes, which is certainly not the case. If the Principle holds, then the molecules in a database that are most similar to a bioactive reference structure are (all other things being equal) those that are most likely to exhibit the reference structure's bioactivity. Ranking the database in order of decreasing similarity, where the similarity is defined using some quantitative measure of inter-molecular similarity, hence provides a rational way of prioritising compounds for biological testing and thus a firm basis for the development of similarity searching methods. It is appropriate to mention here the closely related concept of neighbourhood behaviour [37], which

involves relating absolute differences in bioactivity for pairs of molecules to the dissimilarities for those pairs of molecules. This concept has been used to categorise the effectiveness of molecular descriptors for molecular diversity applications [38–40].

Given the importance of the similar property principle, it is hardly surprising that there have been several attempts to demonstrate its applicability. Perhaps the first detailed study was that reported by Willett and Winterman, which showed that simple fingerprint-based similarities could be used to predict a range of physical, chemical, and biological properties in small QSAR datasets (using a "leave-one-out" prediction approach that is discussed later in this review) [41]. Having demonstrated that similarities in structure mirrored similarities in property, these authors then used differences in the strength of this relationship to compare different types of similarity measure. Specifically, they made the assumption that if the Principle holds for some particular dataset, then the extent of the relationship between structure and property that is obtained using some particular similarity measure provides a basis for evaluating the effectiveness of that measure, and hence for comparing the effectiveness of different types of similarity measure. Analogous results were obtained for their QSAR datasets when they were clustered using a range of hierarchic and non-hierarchic clustering methods [42]. The latter work was extended to much larger datasets in two papers by Brown and Martin [43, 44]. These studies were designed to compare the effectiveness of different clustering methods and different types of fingerprint for selecting structurally diverse database subsets, but their detailed experiments demonstrate clearly the general applicability of the Principle. A later paper by Martin et al. provided a direct evaluation of the Principle using structures that had been tested in over 100 assays at Abbott Laboratories [45]. Whilst noting that there were cases where the Principle did not apply, the principal conclusion was that structurally similar compounds do indeed have similar bioactivities, with the latter increasing as the structural similarity is increased. These studies have been taken further in an interesting study by Steffen et al., who show that the Principle also applies when molecular bioactivities are considered across a range of assays, rather than just a single assay as in the other studies cited here [46].

Further demonstrations of the general validity of the Principle come from two near-contemporaneous studies of the applicability of QSAR models. Thus, Sheridan et al. [47] and He and Jurs [48] showed that the more similar a molecule was to molecules in the training set then the more likely it was that an accurate prediction could be made using the QSAR model that had been derived from that training set. More recently, Bostrom et al. analysed sets of protein–ligand complexes from the Protein Data Bank to demonstrate that molecules that are structurally similar tend to bind to a

biological target in the same way, i.e., in addition to eliciting the same biological response, similar molecules achieve this by means of the same mode of action [49]. Finally, the Principle is attracting further support from work in chemogenomics, with recent studies demonstrating: that molecules with similar 2D fingerprints bind to structurally related biological targets [50, 51]; that molecule-based similarities can suggest novel functional relationships between targets that exhibit little sequence similarity [52, 53]; and that pairs of molecules acting on a common target are more likely to be similar than pairs of molecules that do not share a common target [54].

It should be noted that there are many exceptions to the Principle, a situation that Stahura and Bajorath refer to as the similarity paradox [55]. This is especially the case if attention is focused on the relatively small numbers of structurally related molecules that are commonly encountered in QSAR studies [5, 6, 56], where it is not uncommon for very slight changes in structure to bring about large changes in activity (a phenomenon that has been referred to as an "activity cliff" [57, 58]). However, the similar property principle does provide a highly appropriate basis for similarity searching, where similarities are typically computed for large, or very large, numbers of molecules spanning a huge range of structural classes.

3. Components of a Similarity Measure

Any database searching system must be both efficient (i.e., must involve the use of minimal computing resources, typically time and space) and effective (i.e., must retrieve appropriate items from the database that is being searched). Modern computer hardware and software enable highly efficient similarity searches to be carried out on even the largest chemical databases (at least when using the 2D fingerprint approaches that are considered in this chapter), and we hence focus on the factors that control effectiveness. This is determined by the nature of the measure that is used to compute the degree of resemblance between the reference structure and each of the database structures. A similarity measure has three components: the representation that is used to characterise the molecules that are being compared; the weighting scheme that is used to assign differing degrees of importance to the various components of these representations; and the similarity coefficient that is used to provide a quantitative measure of the degree of structural relatedness between a pair of (possibly weighted) structural representations.

3.1. Representations

Very many techniques are available for representing and encoding the structures of 2D chemical molecules [23, 24, 59] and many of these representations have been used for similarity searching [16, 26, 60]. It is common to divide the many techniques into three broad classes of descriptor: whole molecule (sometimes called 1D) descriptors; descriptors that can be calculated from 2D representations of molecules; and descriptors that can be calculated from 3D representations.

Whole molecule descriptors are single numbers, each of which represents a different property of a molecule such as its molecular weight, the numbers of heteroatoms or rotatable bonds, or a computed physicochemical parameter such as logP. A single 1D descriptor is not usually discriminating enough to allow meaningful comparisons of molecules and a molecule is hence normally represented by several (or many) such descriptors [61, 62]. 2D descriptors include topological indices and substructural descriptors. A topological index is a single number that typically characterises a structure according to its size and shape [63, 64]. There are many such indices: the simplest characterise molecules according to their size, degree of branching, and overall shape, while more complex indices take account of the properties of atoms as well as their connectivities. As with 1D descriptors, multiple different indices are normally combined for similarity searching [65]. Substructure-based descriptors characterise a molecule by the substructural features that it contains, either by the molecule's 2D chemical graph or by its fingerprint. Fingerprints are the focus of this chapter and are hence discussed in more detail below. They have been found to be at least as effective, if not more so, for virtual screening than chemical graphs [66] despite the fact that they provide a much less precise representation of a molecule's structure than does the underlying graph (which contains a full description of the molecule's topology). There is hence some interest in the use of simplified graph representations for virtual screening [67–70], and it is likely that work in this area will be developed further in the future. 3D descriptors are inherently more complex since they need to take account of the fact that many molecules are conformationally flexible (although some successful 3D similarity measures have assumed that a molecule can be represented by a single, low-energy conformation). Similarity measures have been reported that are based on inter-atomic distances [71], molecular surfaces [72], electrostatic fields [73, 74], and molecular shapes [75, 76] inter alia.

This chapter focuses on fingerprint-based similarity searching, and it is hence appropriate to discuss the various types of fingerprint that are available in more detail. Fingerprints enable effective similarity searching, but they were first developed for efficient substructure searching. This involves using a subgraph isomorphism algorithm to check for an exact mapping of the atoms and

bonds in a query substructure onto the atoms and bonds of each database structure [23, 24]. Graph matching algorithms are far too slow to enable interactive substructure searching of large files on their own, and it is hence necessary to use an initial *screening* search. This filters out the great majority of the database structures that do not contain all of the substructural fragments present in the query substructure, with only those few molecules that do contain all of these fragments being passed on for the time-consuming graph-matching stage. The presence or absence of fragments in a query substructure or in a database structure is encoded in a binary vector that is normally referred to as a fingerprint.

There are two main ways of selecting the fragments that are encoded in a fingerprint [23, 24, 77, 78]. In a *dictionary-based* approach, there is a pre-defined list of fragments, with normally one fragment allocated to each position in the bit string. A molecule is checked for the presence of each of the fragments in the dictionary, and a bit set (or not set) when a fragment is present (or absent). The dictionary normally contains several different types of fragment. For example, an *augmented atom* contains a central atom together with its neighbouring atoms and bonds, and an *atom sequence* contains a specific number of connected atoms and their intervening bonds. The effectiveness of the dictionary is maximised if a statistical analysis is carried out of the sorts of molecules that are to be fingerprinted, so as to ensure that the most discriminating fragments are included [79–81]. In a *molecule-based* approach, hashing algorithms are used to allocate multiple fragments to each bit position. Here, a generic fragment type is specified, e.g., a chain of four connected non-hydrogen atoms, and a note made of all fragments of that type that occur in a given molecule. Each fragment is converted to a canonical form and then hashed using several (typically two or three) hashing algorithms to set bits in the fingerprint. The first widely used fingerprint of this sort was that developed by Daylight Chemical Information Systems Inc. (at http://www.daylight.com). This fingerprint encodes atom sequences up to a specified length (typically from two to seven atoms), with each such sequence being hashed using multiple hashing procedures so that each bit is associated with multiple fragments and each fragment with multiple bit positions.

Both the dictionary-based and the molecule-based approaches are represented in the fingerprints encountered in operational chemoinformatics systems. For example, the fingerprints produced by Digital Chemistry (formerly Barnard Chemical Information, at http://www.digitalchemistry.co.uk), by Sunset Molecular (at http://www.sunsetmolecular.com), and by Symyx Technologies (formerly MDL Information Systems at http://www.symyx.com) are dictionary-based, the Daylight fingerprints mentioned previously and the fingerprints produced by Accelrys

(at http://www.accelrys.com) are molecule-based (using linear chains and circular substructures, respectively), and the Unity fingerprints produced by Tripos (at http://www.tripos.com) are based on both approaches.

Most of the fingerprints above were originally developed for efficient substructure searching, and it is perhaps surprising that they have also been found to provide a highly effective, alternative type of database access. There are also fingerprints that have been developed specifically for similarity searching [14, 51, 82–87]. It is noteworthy that many of the newer types of fingerprint describe the atoms not by their elemental types but by their physicochemical characteristics, so as to enable the identification of database structures that have similar properties to the reference structure in a similarity search but that have different sets of atoms. This increases the chances of *scaffold-hopping*, i.e., the identification of novel classes of molecule with the requisite bioactivity [88–91]. We should also note that the discussion here is restricted to fingerprints that encode structural fragments: other types of fingerprint used for similarity searching have involved other types of information such as property information [46, 92, 93] or affinities to panels of proteins [94, 95].

3.2. Weighting Schemes

Most fingerprints are binary in nature, with each bit denoting the presence/absence of a substructural fragment in a molecule. However, the elements of a fingerprint can also contain non-binary information that assigns a weight, or degree of importance, to the corresponding features. Thus, a feature that had a large weight and that occurred in both the reference structure and a database structure would contribute more to the overall similarity of those molecules than would a common feature with a small weight. Weighting features in fingerprints lies at the heart of many approaches to substructural analysis and related machine-learning approaches where large amounts of training data are available (vide infra) [27, 96, 97], but has been much less studied in the context of similarity searching, where the only information that is available is the reference structure and the database structures that are to be searched.

Willett and Winterman suggested that three types of weighting could be used for fingerprint-based similarity searching: weighting based on the number of times that a fragment occurred in an individual molecule; weighting based on the number of times that a fragment occurred in an entire database; and weighting based on the total number of fragments within a molecule [41]. Of these three types of weight, the last is accommodated in many of the common similarity coefficients (vide infra) since they include a factor describing the sizes (in terms of numbers of fragments) of the two molecules that are being compared, whilst studies of the second type of weight have been limited to date

[98, 99]. However, there have been several studies of the use of information about fragment occurrences in a single molecule [41, 43, 70, 84, 85, 100–102]. These studies have suggested that fingerprints encoding the occurrences of substructural fragments may be able to give better screening performance than conventional, binary fingerprints. However, the results have been far from consistent; and the performance differences often quite small; many of the previous studies were limited, either in terms of the numbers of molecules involved or in the extent to which the weighted and binary fingerprints differed; and there has been no attempt to explain the observed levels of performance. This situation has been addressed in a recent study by Arif et al. [103], which has demonstrated conclusively the general superiority of occurrence-based weighting and also rationalised the different (and sometimes very different) levels of performance that were observed in experiments involving a range of weighting schemes, types of fingerprint and chemical databases. Their recommended scheme involves encoding both the reference structure and the database structures using the square root of a fragment's occurrence; the study was, however, limited to the use of the Tanimoto coefficient (vide infra) and it remains to be seen whether analogous results are obtained with other types of coefficient.

3.3. Similarity Coefficients

The calculation of inter-object similarities by means of a similarity coefficient lies at the heart of cluster analysis, a multivariate data analysis technique that is used across the sciences and social sciences [104], and very many different similarity coefficients have thus been developed for this purpose [105, 106]. Willett et al. provide an extended account of those that have been used for applications in chemoinformatics [35], focusing on the mathematical characteristics of the various coefficients that they discuss and, in particular, on the broad class of similarity coefficients known as *association coefficients*. These are all based on the number of fragments, i.e., bits in a fingerprint, common to the fingerprints describing a reference structure and a database structure, with this number normalised by some function based on the numbers of non-zero bits in the two fingerprints that are being compared. An example of an association coefficient is the Tanimoto coefficient. This was found to work well in Willett and Winterman's early similarity study of QSAR datasets [41] and was hence adopted as the coefficient of choice when the first operational searching systems were introduced a few years later. Subsequent work has demonstrated the appropriateness of this choice: the Tanimoto coefficient has been found to perform well in a wide range of applications, and not just similarity searching, and remains the yardstick against which alternative approaches are judged, despite the many years that have passed since Willett and Winterman's initial study in 1986. Like most association

coefficients, the Tanimoto coefficient takes values between zero and unity when used with binary fingerprints: a value of zero corresponds to two fingerprints that have no bits in common, while a value of unity corresponds to two identical fingerprints [35].

Whilst widely used, the Tanimoto coefficient is known to give low similarity values in searches for small reference structures (where just a few bits are switched on in the reference structure's fingerprint) [107–109], and is also known to have an inherent bias towards specific similarity values [110]. These observations spurred several comparative studies (summarised in [28]) that involved over 20 different fingerprint-based similarity coefficients. None of the coefficients was found to be consistently superior to the Tanimoto coefficient, and it was shown (both experimentally and theoretically) that most coefficients exhibit at least some degree of dependence on the sizes (i.e., numbers of set bits) of the molecules that are compared in a similarity search. Later studies have focussed on the use of asymmetric coefficients, based on ideas first put forward by Tversky [111], for the calculation of inter-molecular structural similarities [112, 113]. In a symmetric coefficient, the value of the coefficient is independent of whether a reference structure is mapped to a database structure or vice versa. This is not so with asymmetric coefficients and it has been suggested that this may be beneficial for database searching [30, 114], although the merits of such coefficients are still the subject of debate [115, 116].

The coefficients discussed thus far focus on the substructural fragments that are common to a reference structure and a database structure, i.e., those positions in the fingerprint where the bit is switched on. Information about the other bits, i.e., those that are switched off, may be included implicitly, typically via a contribution to the overall coefficient that reflects sizes of the two molecules that are being compared. Extended versions have been reported of the Tanimoto and Tversky coefficients where the overall value of the coefficient is the weighted sum of one coefficient based on the bits switched on and of one coefficient based on the bits switched off [109, 117].

Association coefficients are specifically designed for use with binary data. If interval or ratio data is used, as would be the case if some form of fragment weighting scheme was to be employed in the generation of a fingerprint, other types of coefficient may then be appropriate. The Euclidean distance has been found to work well in many data analysis studies, both in chemoinformatics and more generally [35, 104]; however, Varin et al. [118] have recently suggested that a coefficient described by Gower and Legendre [119], which reduces to the Tanimoto coefficient when applied to binary data, performs very well when weighted fingerprints are used for clustering and similarity searching.

4. Evaluation of Similarity Measures

It will be clear from the above that there are very many possible combinations of fingerprint, coefficient, and weighting scheme that could be used to build a similarity measure for similarity searching. It is hence reasonable to ask how one can assess the effectiveness of different measures and thus how one can identify the most appropriate for a particular searching application.

The aim of similarity searching, as of any virtual screening method, is to identify bioactive molecules and the evaluation of search effectiveness is hence normally carried out using datasets for which both structure and bioactivity data are available. There is, of course, a vast amount of such data available in corporate databases as a result of the massive biological screening programs carried out by industry, but intellectual property considerations mean that this rarely, or ever, becomes available for more general use. This is a severe limitation since the development of the science of similarity searching requires standard datasets that can be used for the evaluation and comparison of different methods as they become available. Instead, most reported studies of similarity measures make use of a limited number of public datasets for which both structural and activity data are available. Examples of such datasets that have become widely used include the *MDL Drug Data Report* database (available from Symyx Technologies at http://www.symyx.com), the *World of Molecular Bioactivity* database (available from Sunset Molecular at http://sunsetmolecular.com/), the National Cancer Institute AIDS database (available from the National Library of Medicine Developmental Therapeutics Programme at http://dtp.nci.nih.gov), and the *Directory of Useful Decoys* (DUD) database (available from http://www.dud.docking.org/).

Although standard datasets are widely used, it is important to recognise that they do have some limitations. First, they contain molecules that have been reported as exhibiting some particular bioactivity but may say nothing as to their activity or inactivity against other biological targets; instead, it is normally the case that the absence of activity information is taken to mean inactivity. Second, molecules that reach the published literature (and that are hence eligible for inclusion in such databases) may be only a small, carefully studied, and high-quality subset of those that were actually synthesised and tested in a screening program. Third, the "me too" or "fast follower" nature of research in the pharmaceutical industry means that some structural classes are overly represented in a dataset. Finally, the numbers of molecules in these datasets are typically an order of magnitude less than in

corporate databases, which may contain several million molecules. Notwithstanding these characteristics, the existence of these datasets does mean that there is a natural platform for evaluating new methods and for comparing them with existing methods.

The bioactivity data can be either *qualitative* (e.g., a molecule is categorised as either active or inactive) or *quantitative* (e.g., an IC50 value is available for a molecule), but the Similar Property Principle provides the basis for performance evaluation irrespective of the precise nature of the biological data. If the Principle does hold for a particular dataset, i.e., if structurally Similar molecules have similar activities, then the nearest-neighbour molecules in a similarity search are expected to have the same activity as the bioactive reference structure. The effectiveness of a similarity measure can hence be quantified by determining the extent to which the similarities resulting from its use mirror similarities in the bioactivity of interest.

Several reviews are available of effectiveness measures that can be used when qualitative activity data are available [38, 120, 121]. Most if not all of the common measures can be regarded as a function of one or both of two underlying variables: the *recall* and the *precision*. Assume that a similarity search has been carried out, and a threshold applied to the resulting ranked list to retrieve some small subset, e.g., 1%, of the database. Then the recall is the fraction of the total number of active molecules retrieved in that subset; and the precision is the fraction of that subset that is active. A good search is one that maximises both recall and precision so that, in the ideal case, a user would be presented with all of the actives in the database without any additional inactives: needless to say, this ideal is very rarely achieved in practice.

Examples of measures that have been extensively used include the *enrichment factor*, i.e., the number of actives retrieved relative to the number that would have been retrieved if compounds had been picked from the database at random [122], the numbers of actives that have been retrieved at some fixed position in the ranking [123], and the receiver operating characteristic (or ROC curve) [124, 125]. An ROC curve plots the percentage of true positives retrieved against the percentage of false positives retrieved at each position in the ranking (or at some series of fixed positions, e.g., the top 5%, the top 10%, the top 15%, etc.). ROC curves are widely used in machine learning and pattern recognition research, but their use in virtual screening has been criticised [126] since no particular attention is paid to the top-ranked molecules, and it is these that would actually be selected for testing in an operational screening system. There is much current interest in the evaluation of virtual screening (based on similarity searching, docking, or whatever) and it is likely to be

some time before full agreement is reached as to the best approaches to evaluation [127, 128].

Similarity searching is normally used in the lead discovery stage of a drug discovery programme, when only qualitative biological data are available and when the evaluation criteria mentioned in the previous paragraph are appropriate. However, the Similar Property Principle can also be applied to the analysis of datasets with quantitative data, using a leave-one-out approach analogous to those used in QSAR studies [121]. Assume that the activity value for the reference structure R is known and is denoted by $A(R)$. A similarity search is carried out and some number of R's nearest neighbours identified. The predicted activity value for R, $P(R)$, is then taken to be the arithmetic mean of the known activity values for this set of nearest neighbours. The similarity search is repeated using different reference structures, and the correlation coefficient is then computed between the resulting sets of $A(R)$ and $P(R)$ values. A large correlation coefficient implies a good fit between the known and predicted bioactivities and hence strict adherence to the similar property principle by the similarity search procedure that was used to generate the sets of nearest neighbours. This approach to performance evaluation was pioneered by Adamson and Bush [34]; it formed the basis for Willett's extensive studies of similarity and clustering methods in the 1980s [42] and, more recently, was used in Brown and Martin's much-cited comparison of structural descriptors for compound selection [43, 44].

A focus on the number of active molecules retrieved by a similarity search is entirely reasonable, but the needs of lead discovery mean that it is also important to consider the structural diversity of those active molecules [129]. Specifically, account needs to be taken of the scaffold-hopping abilities of the similarity search since, e.g., a search retrieving 25 active analogues that all have the same scaffold as the reference structure is likely to be of much less commercial importance than a search retrieving just five actives if each of these has a different scaffold. It is often suggested that fragment-based 2D similarity searching has only a limited scaffold-hopping capability, especially when compared with more complex (and often much more time-consuming) 3D screening methods. This suggestion is clearly plausible, but there is a fair amount of evidence to suggest that 2D methods can exhibit non-trivial scaffold-hopping capabilities [16].

The evaluation criteria described above have been used in a very large, and constantly increasing, number of studies that discuss the effectiveness of similarity searching. Even a brief discussion of these many studies would require a totally disproportionate amount of space, and the reader is accordingly referred to the many excellent reviews that exist [2, 5–7, 35, 60].

5. Use of Multiple Reference Structures

As discussed thus far, similarity searching has involved matching a single bioactive reference structure against a database using a single similarity measure. Over the last few years, perhaps the principal development in the field of similarity searching has been the appearance of a range of methods that involve the use of additional information in generating a ranking of the database. It is possible to identify two broad classes of approach: the first class involves the use of *data fusion*, or *consensus*, methods; while the second class involves the use of *machine learning* methods to develop predictive models that can guide future searches given a body of training data. It is debatable where similarity searching stops and where machine learning starts, but the main difference is in the amounts of bioactivity data available and the way that data is used. One of the principal attractions of similarity searching as a tool for virtual screening is that it requires just a single known active molecule, whereas the application of machine learning to virtual screening requires a pool of molecules (this pool ideally including not just actives but also inactives) to enable the development of a predictive model. In this review we shall focus more on data fusion since work in this area is more tightly aligned to conventional similarity searching, but make some remarks about machine learning approaches at the end of the section.

The comparative studies referenced in Subsection 4 have typically sought to identify a single "best" similarity method; hardly surprisingly, it has not been possible to identify a single approach that is consistently superior to all others across a range of reference structures, biological targets, and performance criteria [7, 16]. The data fusion approach involves carrying out multiple similarity searches and then combining the resulting search outputs to give a single fused output that is presented to the searcher. For example, assume that three different types of 2D fingerprint are available. A search is carried out using the first fingerprint type to describe the reference structure and each of the database structures, and the database ranked in decreasing order of the computed similarity. The procedure is repeated using each of the other two types of fingerprint in turn, and the three database rankings are then combined using a fusion rule, e.g., taking the mean rank for each database structure when averaged across the three rankings. Data fusion was first used for similarity searching in the mid-1990s as discussed in an extensive review by Willett [130]; analogous techniques are used in docking, where the approach is called *consensus scoring* [131].

Early studies of data fusion involved combining searches that were based on different types of structural representation.

For example, Ginn et al. reported studies involving a wide range of types of representation (2D fingerprints, sets of physicochemical properties, molecular electrostatic potential descriptors, and infrared spectral descriptors) and of combination rules [132, 133]. This work, and analogous studies by the Sheridan group [122, 134], suggested that fusion could give search outputs that were more robust, in the sense of offering a consistently high level of performance, than those obtainable from the use of a single type of similarity search. More recent work in this area has considered the combination of further types of representation, and the combination of searches that involve different similarity coefficients [135, 136].

Thus far, we have considered data fusion to involve a single reference structure but multiple similarity measures, an approach that Whittle et al. refer to as *similarity fusion* [137]. The alternative, *group fusion* approach inverts the relationship between similarity measure and reference structure, so that the multiple searches that are input to the fusion procedure result from using multiple reference structures and a single similarity measure (e.g., the Tanimoto coefficient and 2D fingerprints). This idea seems to have been first reported by Xue et al. [138] and then by Schuffenhauer et al. [51] some time after the initial studies of similarity fusion; however, group fusion appears from the literature to have become much more widely used. Its popularity dates from a study by Hert et al. [123] who found that fusing the similarity rankings obtained from as few as ten reference structures enabled searches to be carried out that were comparable to even the very best from amongst many hundreds of conventional similarity searches using individual reference structures. Subsequent studies demonstrated the general validity of the approach, and it has now been widely adopted [139, 140].

Hert et al. have also described a modification of conventional similarity searching that makes use of group fusion [141, 142]. A similarity search is carried out in the normal way using a single reference structure, and the nearest neighbours identified. The assumption is then made that they also are active, as is likely to be the case if the similar property principle applies to the search. Each of these nearest neighbours is used in turn as a reference structure for a further similarity search, and the complete set of rankings (one from the original reference structure and one from each of the nearest neighbours) is then fused to give the final output ranking. This *turbo similarity searching* approach resulted in searches that were nearly always superior to conventional similarity searching (where just the initial reference structure is used) in its ability to identify active molecules, although performance appears to be crucially dependent on the effectiveness of the initial search based on the original reference structure [143].

Most studies of fusion methods have found that they seem to work well in practice but have not provided any rationale for why this might be so [130]. Two studies have addressed this question. An empirical study by Baber et al. [144] showed that active molecules are more tightly clustered than are inactive molecules (as would indeed be expected if the Similar Property Principle holds). Thus, when multiple scoring functions are used in similarity fusion, they are likely to repeatedly select many actives but not necessarily the same inactives, providing an enrichment of actives at the top of the final fused ranking. Whittle et al. provide a rigorous theoretical approach to the modelling of data fusion [145, 146]. Their model suggests that the origin of performance enhancement for simple fusion rules can be traced to a combination of differences between the retrieved active and retrieved inactive similarity distributions and the geometrical difference between the regions of these multivariate distributions that the chosen fusion rule is able to access. Although their model gave predictions in accord with experimental data, it was concluded that improvements over conventional similarity searching would be obtained only if large amounts of training data are available; however, this is not normally the case in the early stages of drug-discovery programmes where similarity searching is most commonly used.

Group fusion requires multiple reference structures, but the processing involves them being treated on an individual basis, with each one generating their own similarity ranking. It is arguable that this wastes available information since it takes no account of the relationships between the reference structures, as reflected in the bits that are, and that are not, set in their fingerprints. This is valuable information that can be correlated with the other information that we have available, i.e., that these reference structures are known to exhibit the activity that is being sought in the similarity search. Put simply, if a bit is set in many of the reference structures' fingerprints, then it seems likely that the corresponding 2D fragment is positively associated with the activity of interest, and this information can be used to enhance the effectiveness of a similarity search.

The relationship between fragment occurrences and bioactivity in large databases was first studied by Cramer et al. [147]. Their *substructural analysis* approach [148–151] and the closely related *naïve Bayesian classifier* [82, 142, 152–154] are widely used examples of the application of machine learning methods to virtual screening [97]. These applications require considerable amounts of training: this is normally HTS data that contains many examples of both active and inactive molecules. The use of such approaches for similarity searching typically uses training data based on the set of reference structures (for the

actives) and on any large set of molecules from which the known actives have been removed (for the inactives). One example of this approach is the MOLPRINT system of Bender et al. [82, 155], who have used a naïve Bayesian classifier with atom-centred substructures chosen using a feature selection algorithm. However, the largest body of work in this area has been carried out by the Bajorath group, who have used a Bayesian approach to derive functions that relate the probability of a molecule exhibiting bioactivity to the statistical distributions of the descriptor values for that molecule's descriptors [156]. The procedure involves estimating the probability that a molecule will be active given a particular value of a descriptor, where the descriptor can be binary (as with a bit in a fingerprint) or non-binary (as with a molecular property). The probabilities of activity for different descriptors are assumed to be statistically independent, and it is hence possible to compute the overall probability of activity (or inactivity) for a molecule by taking the product of the individual descriptor probabilities. It should be noted that the independence assumption is generally incorrect (indeed, it is naïve, which is why the approach is called a naïve Bayesian classifier) but has been found to work well in practice. The overall approach is markedly more complex than with group fusion, where the reference structures are used for individual similarity searches; however, detailed comparisons suggest the greater search effectiveness of the Bayesian approach [157]. An interesting application of this work is the ability to predict the probability that a similarity approach will be able to identify novel molecules that exhibit the reference structures' bioactivity when searching a particular database: if this probability is low, then it may be worth considering an alternative type of structure representation for the search [156]. Other recent studies by this group have included: ways of weighting the bits in fingerprints [158]; the use of quantitative, rather than qualitative, bioactivities for the training data [159]; and the use of a different machine learning tool, a support vector machine, for similarity searching [160].

We have thus considered two ways of using multiple reference structures: combining rankings based on each structure in turn (group fusion) and combining information about the bits that are and are not set in the structures' fingerprints. There is a much simpler approach, involving the combination of the multiple reference structures' fingerprints into a single, combined fingerprint [51, 161]; however, this appears to be less effective than the other two approaches [123, 162]. There is also a considerably more complex approach, which involves combining the actual chemical graphs of the reference structures (rather than fingerprints derived from those graphs) [163]; however, this hardly comes within the scope of a review of fingerprint-based methods.

6. Conclusions

Similarity searching of chemical databases using 2D structural fingerprints was first described almost a quarter of a century ago. Since that time, it has established itself as one of the most valuable ways of accessing a chemical database to identify novel bioactive molecules, providing a natural complement to the long-established systems for 2D substructure searching. It is now routinely used in the initial stages of virtual screening programmes, where very little structure–activity data may be available at the start of a research project, and has proved to be remarkably effective in this role, despite the inherent simplicity of the methods that are being used. There are very many different types of similarity measure that can be used to determine the similarity between a pair of molecules: at present, the Tanimoto coefficient and binary fingerprints are the method of choice, but it would be surprising if it did not prove possible to identify more effective ways of searching, e.g., using some type of fragment weighting scheme. Current research in similarity searching is looking at ways of exploiting the information that is available when multiple reference structures are available.

References

1. Rouvray, D. H. (1990) The evolution of the concept of molecular similarity, in *Concepts and Applications of Molecular Similarity* (Johnson, M. A., and Maggiora, G. M., Eds.), pp 15–42, John Wiley, Chichester.
2. Bender, A., and Glen, R. C. (2004) Molecular similarity: a key technique in molecular informatics. *Organic and Biomolecular Chemistry* **2**, 3204–3218.
3. Dean, P. M., (Ed.) (1994) *Molecular Similarity in Drug Design*, Chapman and Hall, Glasgow.
4. Downs, G. M., and Willett, P. (1995) Similarity searching in databases of chemical structures. *Reviews in Computational Chemistry* **7**, 1–66.
5. Maldonado, A. G., Doucet, J. P., Petitjean, M., and Fan, B.-T. (2006) Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular Diversity* **10**, 39–79.
6. Nikolova, N., and Jaworska, J. (2003) Approaches to measure chemical similarity – a review. *Quantitative Structure-Activity Relationships and Combinatorial Science* **22**, 1006–1026.
7. Sheridan, R. P., and Kearsley, S. K. (2002) Why do we need so many chemical similarity search methods? *Drug Discovery Today* **7**, 903–911.
8. Alvarez, J., and Shoichet, B., (Eds.) (2005) *Virtual Screening in Drug Discovery*, CRC Press, Boca Raton.
9. Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery* **1**, 882–894.
10. Böhm, H.-J., and Schneider, G., (Eds.) (2000) *Virtual Screening for Bioactive Molecules,* Wiley-VCH, Weinheim.
11. Klebe, G., (Ed.) (2000) *Virtual Screening: An Alternative or Complement to High Throughput Screening*, Kluwer, Dordrecht.
12. Lengauer, T., Lemmen, C., Rarey, M., and Zimmermann, M. (2004) Novel technologies for virtual screening. *Drug Discovery Today* **9**, 27–34.
13. Oprea, T. I., and Matter, H. (2004) Integrating virtual screening in lead discovery. *Current Opinion in Chemical Biology* **8**, 349–358.
14. Gedeck, P., Rhode, B., and Bartels, C. (2006) QSAR – how good is it in practice?

Comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling* **46**, 1924–1936.
15. McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.-F., and Cornell, W. D. (2007) Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling* **47**, 1504–1519.
16. Sheridan, R. P. (2007) Chemical similarity searches: when is complexity justified? *Expert Opinion on Drug Discovery* **2**, 423–430.
17. Sheridan, R. P., McGaughey, G. B., and Cornell, W. D. (2008) Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *Journal of Computer-Aided Molecular Design* **22**, 257–265.
18. Talevi, A., Gavernet, L., and Bruno-Blanch, L. E. (2009) Combined virtual screening strategies. *Current Computer-Aided Drug Design* **5**, 23–37.
19. Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., Lindvall, M., Nevins, N., Semus, S. F., Senger, S., Tedesco, G., Wall, I. D., Woolven, J. M., Peishoff, C. E., and Head, M. S. (2006) A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry* **49**, 5912–5931.
20. Wilton, D., Willett, P., Lawson, K., and Mullier, G. (2003) Comparison of ranking methods for virtual screening in lead-discovery programs. *Journal of Chemical Information and Computer Sciences* **43**, 469–474.
21. Bajorath, J., (Ed.) (2004) *Chemoinformatics Concepts, Methods and Tools for Drug Discovery*, Humana Press, Totowa NJ.
22. Gasteiger, J., and Engel, T., (Eds.) (2003) *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim.
23. Leach, A. R., and Gillet, V. J. (2007) *An Introduction to Chemoinformatics*, 2nd edition, Kluwer, Dordrecht.
24. Gasteiger, J., (Ed.) (2003) *Handbook of Chemoinformatics*, Wiley-VCH, Weinheim.
25. Johnson, M. A., and Maggiora, G. M., (Eds.) (1990) *Concepts and Applications of Molecular Similarity*. John Wiley, New York.
26. Willett, P. (2009) Similarity methods in chemoinformatics. *Annual Review of Information Science and Technology* **43**, 3–71.
27. Eckert, H., and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches. *Drug Discovery Today* **12**, 225–233.
28. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **11**, 1046–1053.
29. Hagadone, T. R. (1992) Molecular substructure similarity searching – efficient retrieval in two-dimensional structure databases. *Journal of Chemical Information and Computer Sciences* **32**, 515–521.
30. Senger, S. (2009) Using Tversky similarity searches for core hopping: finding the needles in the haystack. *Journal of Chemical Information and Modeling* **49**, 1514–1524.
31. Willett, P. (1985) An algorithm for chemical superstructure searching. *Journal of Chemical Information and Computer Sciences* **25**, 114–116.
32. Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom pairs as molecular-features in structure activity studies – definition and applications. *Journal of Chemical Information and Computer Sciences* **25**, 64–73.
33. Willett, P., Winterman, V., and Bawden, D. (1986) Implementation of nearest-neighbour searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences* **26**, 36–41.
34. Adamson, G. W., and Bush, J. A. (1973) A method for the automatic classification of chemical structures. *Information Storage and Retrieval* **9**, 561–568.
35. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching. *Journal of Chemical Information and Computer Sciences* **38**, 983–996.
36. Wilkins, C. L., and Randic, M. (1980) A graph theoretical approach to structure-property and structure-activity correlation. *Theoretica Chimica Acta* **58**, 45–68.
37. Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. (1996) Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors. *Journal of Medicinal Chemistry* **39**, 3049–3059.
38. Dixon, S. L., and Merz, K. M. (2001) One-dimensional molecular representations and similarity calculations: methodology and validation. *Journal of Medicinal Chemistry* **44**, 3795–3809.
39. Papadatos, G., Cooper, A. W. J., Kadirkamanathan, V., Macdonald, S. J. F., McLay, I. M., Pickett, S. D., Pritchard, J. M., Willett, P., and Gillet, V. J. (2009) Analysis of neighborhood behaviour in lead optimisation and array design. *Journal of Chemical Information and Modeling* **49**, 195–208.

40. Perekhodtsev, G. D. (2007) Neighbourhood behavior: validation of two-dimensional molecular similarity as a predictor of similar biological activities and docking scores. *QSAR and Combinatorial Science* **26**, 346–351.
41. Willett, P., and Winterman, V. (1986) A comparison of some measures of intermolecular structural similarity. *Quantitative Structure-Activity Relationships* **5**, 18–25.
42. Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems*, Research Studies Press, Letchworth.
43. Brown, R. D., and Martin, Y. C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *Journal of Chemical Information and Computer Sciences* **36**, 572–584.
44. Brown, R. D., and Martin, Y. C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *Journal of Chemical Information and Computer Sciences* **37**, 1–9.
45. Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002) Do structurally similar molecules have similar biological activities? *Journal of Medicinal Chemistry* **45**, 4350–4358.
46. Steffen, A., Kogej, T., Tyrchan, C., and Engkvist, O. (2009) Comparison of molecular fingerprint methods on the basis of biological profile data. *Journal of Chemical Information and Modeling* **49**, 338–347.
47. Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences* **44**, 1912–1928.
48. He, L., and Jurs, P. C. (2005) Assessing the reliability of a QSAR model's predictions. *Journal of Molecular Graphics and Modelling* **23**, 503–523.
49. Bostrom, J., Hogner, A., and Schmitt, S. (2006) Do structurally similar ligands bind in a similar fashion? *Journal of Medicinal Chemistry* **49**, 6716–6725.
50. Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006) Global mapping of pharmacological space. *Nature Biotechnology* **24**, 805–815.
51. Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of Chemical Information and Computer Sciences* **43**, 391–405.
52. Hert, J., Keiser, M. J., Irwin, J. J., Oprea, T. I., and Shoichet, B. K. (2008) Quantifying the relationship among drug classes. *Journal of Chemical Information and Modeling* **48**, 755–765.
53. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nature Biotechnology* **25**, 197–206.
54. Cleves, A. E., and Jain, A. N. (2006) Robust ligand-based modeling of the biological targets of known drugs. *Journal of Medicinal Chemistry* **49**, 2921–2938.
55. Stahura, F. L., and Bajorath, J. (2002) Bio- and chemo-informatics beyond data management: crucial challenges and future opportunities. *Drug Discovery Today* **7**, S41–S47.
56. Kubinyi, H. (1998) Similarity and dissimilarity: a medicinal chemist's view. *Perspectives in Drug Discovery and Design* **9–11**, 225–232.
57. Maggiora, G. M. (2006) On outliers and activity cliffs – why QSAR often disappoints. *Journal of Chemical Information and Modeling* **46**, 1535.
58. Peltason, L., and Bajorath, J. (2007) SAR index: quantifying the nature of structure-activity relationships. *Journal of Medicinal Chemistry* **50**, 5571–5578.
59. Todeschini, R., and Consonni, V. (2002) *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim.
60. Glen, R. C., and Adams, S. E. (2006) Similarity metrics and descriptor spaces – which combinations to choose? *QSAR and Combinatorial Science* **25**, 1133–1142.
61. Godden, J. W., Xue, L., Kitchen, D. B., Stahura, F. L., Schermerhorn, E. J., and Bajorath, J. (2002) Median partitioning: a novel method for the selection of representative subsets from large compound pools. *Journal of Chemical Information and Computer Sciences* **42**, 885–893.
62. Godden, J. W., Furr, J. R., Xue, L., Stahura, F. L., and Bajorath, J. (2004) Molecular similarity analysis and virtual screening by mapping of consensus positions in binary-tansformed chemical descriptor spaces with variable dimensionality. *Journal of Chemical Information and Computer Sciences* **44**, 21–29.
63. Kier, L. B., and Hall, H. L. (1986) *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York.
64. Lowell, H., Hall, H. L., and Kier, L. B. (2001) Issues in representation of molecular structure: the development of molecular connectivity. *Journal of Molecular Graphics and Modelling* **20**, 4–18.

65. Estrada, E., and Uriarte, E. (2001) Recent advances on the use of topological indices in drug discovery research. *Current Medicinal Chemistry* **8**, 1573–1588.
66. Raymond, J. W., and Willett, P. (2002) Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *Journal of Computer-Aided Molecular Design* **16**, 59–71.
67. Rarey, M., and Dixon, J. S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design* **12**, 471–490.
68. Rarey, M., and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design* **15**, 497–520.
69. Barker, E. J., Buttar, D., Cosgrove, D. A., Gardiner, E. J., Gillet, V. J., Kitts, P., and Willett, P. (2006) Scaffold-hopping using clique detection applied to reduced graphs. *Journal of Chemical Information and Modeling* **46**, 503–511.
70. Stiefl, N., Watson, I. A., Baumann, K., and Zaliani, A. (2006) ErG: 2D pharmacophore descriptions for scaffold hopping. *Journal of Chemical Information and Modeling* **46**, 208–220.
71. Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., and Labaudiniere, R. F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry* **42**, 3251–3264.
72. Mount, J., Ruppert, J., Welch, W., and Jain, A. N. (1999) Icepick: a flexible surface-based system for molecular diversity. *Journal of Medicinal Chemistry* **42**, 60–66.
73. Cheeseright, T., Mackey, M., Rose, S., and Vinter, A. (2006) Molecular field extrema as descriptors of biological activity: definition and validation. *Journal of Chemical Information and Modeling* **46**, 6650–6676.
74. Mestres, J., Rohrer, D. C., and Maggiora, G. M. (1997) MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches. *Journal of Computational Chemistry* **18**, 934–954.
75. Ballester, P. J., and Richards, W. G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *Journal of Computational Chemistry* **28**, 1711–1723.
76. Rush, T. S., Grant, J. A., Mosyak, L., and Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *Journal of Medicinal Chemistry* **48**, 1489–1495.
77. Barnard, J. M. (1993) Substructure searching methods – old and new. *Journal of Chemical Information and Computer Sciences* **33**, 532–538.
78. Brown, N. (2009) Chemoinformatics – an introduction for computer scientists. *ACM Computing Surveys*.
79. Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M. (1973) Strategic considerations in the design of screening systems for substructure searches of chemical structure files. *Journal of Chemical Documentation* **13**, 153–157.
80. Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002) Re-optimisation of MDL keys for use in drug discovery. *Journal of Chemical Information and Modeling* **42**, 1273–1280.
81. Hodes, L. (1976) Selection of descriptors according to discrimination and redundancy – application to chemical-structure searching. *Journal of Chemical Information and Computer Sciences* **16**, 88–93.
82. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Molecular similarity searching using atom environments: information-based feature selection and a naive Bayesian classifier. *Journal of Chemical Information and Computer Sciences* **44**, 170–178.
83. Bender, A., Jenkins, J. L., Scheiber, J., Sukuru, S. C. K., Glick, M., and Davies, J. W. (2009) How similar are similarity searching methods? A principal components analysis of molecular descriptor space. *Journal of Chemical Information and Modeling* **49**, 108–119.
84. Ewing, T. J. A., Baber, J. C., and Feher, F. (2006) Novel 2D fingerprints for ligand-based virtual screening. *Journal of Chemical Information and Modeling* **46**, 2423–2431.
85. Fechner, U., Paetz, J., and Schneider, G. (2005) Comparison of three holographic fingerprint descriptors and their binary counterparts. *QSAR and Combinatorial Science* **24**, 961–967.
86. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Organic and Biomolecular Chemistry* **2**, 3256–3266.
87. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-hopping" by topological pharmacophore search: a

contribution to virtual screening. *Angewandte Chemie-International Edition* **38**, 2894–2896.
88. Böhm, H.-J., Flohr, A., and Stahl, M. (2004) Scaffold hopping. *Drug Discovery Today: Technologies* **1**, 217–224.
89. Brown, N., and Jacoby, E. (2006) On scaffolds and hopping in medicinal chemistry. *Mini-Reviews in Medicinal Chemistry* **6**, 1217–1229.
90. Schneider, G., Schneider, P., and Renner, S. (2006) Scaffold-hopping: how far can you jump? *QSAR and Combinatorial Science* **25**, 1162–1171.
91. Martin, Y. C., and Muchmore, S. (2009) Beyond QSAR: lead hopping to different structures. *QSAR & Combinatorial Science* **28**, 797–801.
92. Eckert, H., and Bajorath, J. (2006) Determination and mapping of activity-specific descriptor value ranges for the identification of active compounds. *Journal of Medicinal Chemistry* **49**, 2284–2293.
93. Xue, L., Godden, J. W., Stahura, F. L., and Bajorath, J. (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *Journal of Chemical Information and Computer Sciences* **43**, 1151–1157.
94. Briem, H., and Lessel, U. F. (2000) In vitro and in silico affinity fingerprints: finding similarities beyond structural classes. *Perspectives in Drug Discovery and Design* **20**, 231–244.
95. Kauvar, L. M., Higgins, D. L., Villar, H. O., Sportsman, J. R., Engqvist-Goldstein, A., Bukar, R., Bauer, K. E., Dilley, H., and Rocke, D. M. (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chemistry & Biology* **2**, 107–118.
96. Ormerod, A., Willett, P., and Bawden, D. (1989) Comparison of fragment weighting schemes for substructural analysis, *Quantitative Structure-Activity Relationships* **8**, 115–129.
97. Goldman, B. B., and Walters, W. P. (2006) Machine learning in computational chemistry. *Annual Reports in Computational Chemistry* **2**, 127–140.
98. Moock, T. E., Grier, D. L., Hounshell, W. D., Grethe, G., Cronin, K., Nourse, J. G., and Theodosiou, J. (1988) Similarity searching in the organic reaction domain. *Tetrahedron Computer Methodology* **1**, 117–128.
99. Downs, G. M., Poirrette, A. R., Walsh, P., and Willett, P. (1993) Evaluation of similarity searching methods using activity and toxicity data, in *Chemical Structures 2. The International Language of Chemistry.* (Warr, W. A., Ed.), pp 409–421, Springer Verlag, Berlin.
100. Azencott, C.-A., Ksikes, A., Swamidass, S. J., Chen, J. H., Ralaivola, L., and Baldi, P. (2007) One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties. *Journal of Chemical Information and Modeling* **47**, 965–974.
101. Chen, X., and Reynolds, C. H. (2002) Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *Journal of Chemical Information and Computer Sciences* **42**, 1407–1414.
102. Olah, M., Bologa, C., and Oprea, T. I. (2004) An automated PLS search for biologically relevant QSAR descriptors. *Journal of Computer-Aided Molecular Design* **18**, 437–449.
103. Arif, S. M., Holliday, J. D., and Willett, P. (2009) Analysis and use of fragment occurrence data in similarity-based virtual screening. *Journal of Computer-Aided Molecular Design* **23**, 655–668.
104. Everitt, B. S., Landau, S., and Leese, M. (2001) *Cluster Analysis*, 4th edition, Edward Arnold, London.
105. Gower, J. C. (1982) Measures of similarity, dissimilarity and distance, in *Encyclopaedia of Statistical Sciences* (Kotz, S., Johnson, N. L., and Read, C. B., Eds.), pp 397–405, John Wiley, Chichester.
106. Hubálek, Z. (1982) Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews of the Cambridge Philosophical Society* **57**, 669–689.
107. Flower, D. R. (1988) On the properties of bit string based measures of chemical similarity. *Journal of Chemical Information and Computer Sciences* **38**, 379–386.
108. Dixon, S. L., and Koehler, R. T. (1999) The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. *Journal of Medicinal Chemistry* **42**, 2887–2900.
109. Fligner, M. A., Verducci, J. S., and Blower, P. E. (2002) A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **44**, 110–119.
110. Godden, J. W., Xue, L., and Bajorath, J. (2000) Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *Journal of Chemical Information and Computer Sciences* **40**, 163–166.

111. Tversky, A. (1977) Features of similarity. *Psychological Review* **84**, 327–352.
112. Bradshaw, J. (1997) Introduction to Tversky similarity measure, in *MUG '97 – 11th Annual Daylight User Group Meeting* Laguna Beach CA.
113. Maggiora, G. M., Mestres, J., Hagadone, T. R., and Lajiness, M. S. (1997) Asymmetric similarity and molecular diversity, in *213th National Meeting of the American Chemical Society, April 13–17, 1997*, San Francisco, CA.
114. Chen, X., and Brown, F. K. (2006) Asymmetry of chemical similarity. *ChemMedChem* **2**, 180–182.
115. Wang, Y., Eckert, H., and Bajorath, J. (2007) Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem* **2**, 1037–1042.
116. Wang, Y., and Bajorath, J. (2008) Balancing the influence of molecular complexity on fingerprint similarity searching. *Journal of Chemical Information and Modeling* **48**, 75–84.
117. Wang, Y., and Bajorath, J. (2009) Development of a compound-class directed similarity coefficient that accounts for molecular complexity effects in fingerprint searching. *Journal of Chemical Information and Modeling* **49**, 1369–1376.
118. Varin, T., Bureau, R., Mueller, C., and Willett, P. (2009) Clustering files of chemical structures using the Székely-Rizzo generalisation of Ward's method. *Journal of Molecular Graphics and Modelling* **28**, 187–195.
119. Gower, J. C., and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification* **5**, 5–48.
120. Edgar, S. J., Holliday, J. D., and Willett, P. (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. *Journal of Molecular Graphics and Modelling* **18**, 343–357.
121. Willett, P. (2004) The evaluation of molecular similarity and molecular diversity methods using biological activity data. *Methods in Molecular Biology* **275**, 51–63.
122. Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T., and Sheridan, R. P. (1996) Chemical similarity using physicochemical property descriptors. *Journal of Chemical Information and Computer Sciences* **36**, 118–127.
123. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of Chemical Information and Computer Sciences* **44**, 1177–1185.
124. Cuissart, B., Touffet, F., Crémilleux, B., Bureau, R., and Rault, S. (2002) The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *Journal of Chemical Information and Computer Sciences* **42**, 1043–1052.
125. Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., and Bertrand, H.-O. (2005) Virtual screening workflow development guided by the "Receiver Operating Characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor type 4. *Journal of Medicinal Chemistry* **48**, 2534–2547.
126. Truchon, J.-F., and Bayly, C. I. (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling* **47**, 488–508.
127. Jain, A. N., and Nicholls, A. (2008) Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design* **22**, 133–139.
128. Nicholls, A. (2008) What do we know and when do we know it? *Journal of Computer-Aided Molecular Design* **22**, 239–255.
129. Good, A. C., Hermsmeier, M. A., and Hindle, S. A. (2004) Measuring CAMD technique performance: a virtual screening case study in the design of validation experiments. *Journal of Computer-Aided Molecular Design* **18**, 529–536.
130. Willett, P. (2006) Data fusion in ligand-based virtual screening. *QSAR and Combinatorial Science* **25**, 1143–1152.
131. Feher, M. (2006) Consensus scoring for protein-ligand interactions. *Drug Discovery Today* **11**, 421–428.
132. Ginn, C. M. R., Turner, D. B., Willett, P., Ferguson, A. M., and Heritage, T. W. (1997) Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *Journal of Chemical Information and Computer Sciences* **37**, 23–37.
133. Ginn, C. M. R., Willett, P., and Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion. *Perspectives in Drug Discovery and Design* **20**, 1–16.
134. Sheridan, R. P., Miller, M. D., Underwood, D. J., and Kearsley, S. K. (1996) Chemical similarity using geometric atom pair descriptors. *Journal of Chemical Information and Computer Sciences* **36**, 128–136.

135. Holliday, J. D., Hu, C.-Y., and Willett, P. (2002) Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry and High-Throughput Screening* **5**, 155–166.
136. Salim, N., Holliday, J. D., and Willett, P. (2003) Combination of fingerprint-based similarity coefficients using data fusion. *Journal of Chemical Information and Computer Sciences* **43**, 435–442.
137. Whittle, M., Gillet, V. J., Willett, P., Alex, A., and Loesel, J. (2004) Enhancing the effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients. *Journal of Chemical Information and Computer Sciences* **44**, 1840–1848.
138. Xue, L., Stahura, F. L., Godden, J. W., and Bajorath, J. (2001) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *Journal of Chemical Information and Computer Sciences* **41**, 746–753.
139. Williams, C. (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Molecular Diversity* **10**, 311–332.
140. Zhang, Q., and Muegge, I. (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *Journal of Medicinal Chemistry* **49**, 1536–1548.
141. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2005) Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information. *Journal of Medicinal Chemistry* **48**, 7049–7054.
142. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2006) New methods for ligand-based virtual screening: use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching. *Journal of Chemical Information and Modeling* **46**, 462–470.
143. Gardiner, E. J., Gillet, V. J., Haranczyk, M., Hert, J., Holliday, J. D., Malim, N., Patel, Y., and Willett, P. (2009) Turbo similarity searching: effect of fingerprint and dataset on virtual-screening performance. *Statistical Analysis and Data Mining* **2**, 103–114.
144. Baber, J. C., Shirley, W. A., Gao, Y., and Feher, M. (2006) The use of consensus scoring in ligand-based virtual screening. *Journal of Chemical Information and Modelling* **46**, 277–288.
145. Whittle, M., Gillet, V. J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: theoretical model. *Journal of Chemical Information and Modeling* **46**, 2193–2205.
146. Whittle, M., Gillet, V. J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: similarity and group fusion. *Journal of Chemical Information and Modeling* **46**, 2206–2219.
147. Cramer, R. D., Redl, G., and Berkoff, C. E. (1974) Substructural analysis. A novel approach to the problem of drug design. *Journal of Medicinal Chemistry* **17**, 533–535.
148. Capelli, A. M., Feriani, A., Tedesco, G., and Pozzan, A. (2006) Generation of a focused set of GSK compounds biased toward ligand-gated ion-channel ligands. *Journal of Chemical Information and Modeling* **46**, 659–664.
149. Cosgrove, D. A., and Willett, P. (1998) SLASH: a program for analysing the functional groups in molecules. *Journal of Molecular Graphics and Modelling* **16**, 19–32.
150. Medina-Franco, J. L., Petit, J., and Maggiora, G. M. (2006) Hierarchical strategy for identifying active chemotype classes in compound databases. *Chemical Biology & Drug Design* **67**, 395–408.
151. Schreyer, S. K., Parker, C. N., and Maggiora, G. M. (2004) Data shaving: a focused screening approach. *Journal of Chemical Information and Computer Sciences* **44**, 470–479.
152. Hassan, M., Brown, R. D., Varma-O'Brien, S., and Rogers, D. (2006) Cheminformatics analysis and learning in a data pipelining environment. *Molecular Diversity* **10**, 283–299.
153. Rogers, D., Brown, R. D., and Hahn, M. (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *Journal of Biomolecular Screening* **10**, 682–686.
154. Xia, X. Y., Maliski, E. G., Gallant, P., and Rogers, D. (2004) Classification of kinase inhibitors using a Bayesian model. *Journal of Medicinal Chemistry* **47**, 4463–4470.
155. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors: evaluation of performance. *Journal of Chemical Information and Computer Sciences* **44**, 1708–1718.
156. Vogt, M., Nisius, B., and Bajorath, J. (2009) Predicting the similarity search performance of fingerprints and their combination with molecular property descriptors using probabilistic and information theoretic modeling.

Statistical Analysis and Data Mining **2**, 123–134.

157. Vogt, M., and Bajorath, J. (2008) Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and molecular fingerprints. *Chemical and Biological Drug Design* **71**, 8–14.
158. Wang, Y., and Bajorath, J. (2008) Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. *Journal of Chemical Information and Modeling* **48**, 1754–1759.
159. Vogt, I., and Bajorath, J. (2007) Analysis of a high-throughput screening data set using potency-scaled molecular similarity algorithms. *Journal of Chemical Information and Modeling* **47**, 367–375.
160. Geppert, H., Horvath, T., Gartner, T., Wrobel, S., and Bajorath, J. (2008) Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *Journal of Chemical Information and Modeling* **48**, 742–746.
161. Shemetulskis, N. E., Weininger, D., Blankey, C. J., Yang, J. J., and Humblet, C. (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets. *Journal of Chemical Information and Computer Sciences* **36**, 862–871.
162. Tovar, A., Eckert, H., and Bajorath, J. (2007) Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem* **2**, 208–217.
163. Hessler, G., Zimmermann, M., Matter, H., Evers, A., Naumann, T., Lengauer, T., and Rarey, M. (2005) Multiple-ligand-based virtual screening: methods and applications of the MTree approach. *Journal of Medicinal Chemistry* **48**, 6575–6584.

Chapter 6

Predicting the Performance of Fingerprint Similarity Searching

Martin Vogt and Jürgen Bajorath

Abstract

Fingerprints are bit string representations of molecular structure that typically encode structural fragments, topological features, or pharmacophore patterns. Various fingerprint designs are utilized in virtual screening and their search performance essentially depends on three parameters: the nature of the fingerprint, the active compounds serving as reference molecules, and the composition of the screening database. It is of considerable interest and practical relevance to predict the performance of fingerprint similarity searching. A quantitative assessment of the potential that a fingerprint search might successfully retrieve active compounds, if available in the screening database, would substantially help to select the type of fingerprint most suitable for a given search problem. The method presented herein utilizes concepts from information theory to relate the fingerprint feature distributions of reference compounds to screening libraries. If these feature distributions do not sufficiently differ, active database compounds that are similar to reference molecules cannot be retrieved because they disappear in the "background." By quantifying the difference in feature distribution using the Kullback–Leibler divergence and relating the divergence to compound recovery rates obtained for different benchmark classes, fingerprint search performance can be quantitatively predicted.

Key words: Bayesian statistics, Compound activity classes, Fingerprints, Information theory, Kullback–Leibler divergence, Prediction of compound recall, Search performance, Virtual screening

1. Introduction

Similarity searching using fingerprints is among the most widely applied methods for ligand-based virtual screening [1–4]. Fingerprints are bit string representations where each bit position typically accounts for the presence or absence of a structural feature, fragment, property, or pharmacophore [5–13]. In addition, fingerprints might also be hashed or folded so that individual positions do not correspond to individual features. Instead, features are mapped to overlapping bit segments [5, 8]. A similarity search involves pairwise comparisons of fingerprints calculated

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_6, © Springer Science+Business Media, LLC 2011

for reference and database compounds and results in a ranking of the database in the order of decreasing similarity to reference molecule(s). These comparison calculations involve the assessment of the similarity of two fingerprints. A variety of different similarity coefficients exist to quantify fingerprint overlap. One of the most popular metrics is the Jaccard or Tanimoto coefficient (Tc) [1, 4].

In its most basic approach, similarity searches involve the calculation of the similarity coefficients for a single reference molecule and each database compound. Historically, this has been one of the first approaches to ligand-based virtual screening and continues to be a popular method. Similarity searching conceptually relies on the similarity property principle [14] stating that similar compounds share similar biological activity. Although this represents a simplistic view of structure–activity relationships, the popularity of similarity searching [1, 2] shows that this principle has its merits.

Frequently, activity information for multiple ligands of a target is available and a straightforward approach to incorporate this information into the search for novel active compounds is to combine the results from multiple individual similarity searches. This might be achieved either at the level of similarity scores as in nearest-neighbor searching [15, 16] or based on the rankings of separate search calculations for each reference molecule. Data fusion methods for combining individual calculations have been explored in detail by Whittle et al. [17, 18] and Hert et al. [19]. Moreover, with the availability of activity information for larger sets of compounds, ligand-based virtual screening can also be carried out using machine learning approaches where "training sets" are utilized that are assumed to reflect the characteristics of active compounds.

Machine learning can also be applied to determine which fingerprint features are essential in order to distinguish active from inactive compounds. However, due to the combinatorial number of possible combinations of features and the often limited number of available reference compounds, there usually is insufficient data to determine feature combinations that are relevant for a given biological activity. Instead, as an approximation, a simpler approach is to assume that features (or bit positions) contribute independently to activity. Thus, for each feature, it is separately evaluated how important it is for activity and then these individual likelihoods are combined to yield an overall estimate of the probability of activity. The rather pragmatic assumption of feature independence is approximate at best but has frequently been shown to yield meaningful results in practical applications [20, 21]. This principle forms the basis of naïve Bayesian classification. The application of the Bayesian framework and the independence assumption to fingerprints result in a "weighting scheme" for

fingerprint positions that ultimately assesses the likelihood of activity for a given compound [22].

The performance of similarity searching generally depends on the compound activity class under study [23, 24] and the molecular representations that are used [24]. In this context, the question arises to what extent a given fingerprint design captures the activity-relevant properties of ligands. Answering this question is of fundamental importance for the selection of fingerprints for similarity searching.

The statistical approach presented herein provides a quantitative estimate for the ability of a similarity search calculation with a particular fingerprint to retrieve active compounds from the screening database [25]. It is based on the calculation of fingerprint feature frequency distributions in reference and database compounds. Measures from information theory can be applied to quantify differences between these frequency distributions. The Kullback–Leibler divergence measure employed here is related to the Bayesian approach [26, 27] and applicable when Bayesian weights are assigned to fingerprint features. This measure of feature divergence is found to correlate well with expected recall performance in similarity searching. Therefore, a linear regression model can be derived that relates the divergence to the fingerprint search performance for a variety of activity classes.

2. Methods

First, the theoretical framework of the methodology is presented. Then, we provide step-by-step instructions for the application of the method. Finally, an application example is given.

2.1. Theory

Statistics and information theory provide the basis of the methodology. First, we introduce a Bayesian approach to virtual screening, which results in a specific fingerprint-based weighting scheme. Then we describe a measure of descriptor divergence that is applicable to fingerprint feature distributions and correlates with the expected performance of fingerprint similarity searching (see also Note 1). The Bayesian approach relies on the estimation of probability distributions for active and inactive compounds. Relating these distributions to each other yields an activity class-specific weighting scheme for fingerprint features.

For multiple reference molecules, fingerprint similarity measures can be combined by using data fusion methods [15–19]. Popular methods include k nearest neighbor (k-NN) calculations where the final similarity score is calculated as the average of the k nearest neighbor scores [15, 16].

These approaches do not take the frequencies of database features into account. However, this can be of critical importance for similarity searching. For instance, for a feature present in 70% of reference molecules, one would assume that it might be a good indicator of activity. However, if this feature is also occurring in 90% of all database compounds then, in fact, the probability of activity for a database compound is about 3.8 times higher if the feature is absent. Thus, weighting bit positions according to their relative frequency is clearly relevant for the assessment of potential activity.

Bit frequency calculations have been described in the context of substructural analysis by Ormerod et al. [22] and Hert et al. [15] and in the context of Bayesian methods by Bender et al. [10]. The basic assumption underlying the assessment of the divergence of feature distributions of reference and database compounds is that bit positions of a fingerprint are independent of each other. As stated above, this assumption is also generally made in naïve Bayesian classification. The independence assumption makes it possible to derive individual weights for each bit position.

A fingerprint $\mathbf{v} = (v_i)_{i=1\ldots n}$ is a binary vector of length n. Each bit position v_i can be viewed as a Bernoulli-distributed random variable whose success probability p_i can be estimated by counting the relative frequency with which a bit is set on in a training set consisting of active and inactive/database compounds (see also Note 2). We will denote the probability of bit position v_i being set on for active compounds with $P(v_i = 1|A) = {p_i}^A$, ${q_i}^A = 1 - {p_i}^A$ and for inactive compounds with $P(v_i = 1|B) = {p_i}^B$, ${q_i}^B = 1 - {p_i}^B$. The likelihood of activity can then be given as the ratio of these two probabilities. On the basis of Bayes' theorem

$$P(A|v_i) = \frac{P(v_i|A)P(A)}{P(v_i)} \tag{1}$$

this ratio is expressed as:

$$\frac{P(A|v_i)}{P(B|v_i)} = \frac{P(v_i|A)P(A)}{P(v_i|B)P(B)} \tag{2}$$

The prior probabilities $P(A)$ and $P(B)$ are in general unknown. When compounds are ranked according to the probability ratio, these prior probabilities only contribute a constant factor. Thus, omitting them does not affect the compound prioritization and hence it is sufficient to consider the ratio

$$R(v_i) = \frac{P_i(v_i|A)}{P_i(v_i|B)} = \left\{ \begin{array}{lll} p_i^A/p_i^B & if & v_i = 1 \\ q_i^A/q_i^B & if & v_i = 0 \end{array} \right\} \tag{3}$$

In order to calculate the overall likelihood for all features $\mathbf{v} = (v_i)_{i=1\ldots n}$ the $R(v_i)$ can be multiplied under the independence

assumption. Introducing logarithms converts the product to a summation:

$$\log R(\mathbf{v}) = \sum_{i=1}^{n} \log R(v_i) = \sum_{i=1}^{n} \log \frac{P(v_i|A)}{P(v_i|B)} \tag{4}$$

For the summation, all possible fingerprint features need to be considered regardless of whether they are set on or off. For fingerprints of relatively small size, this does not pose a problem. For larger fingerprints, especially combinatorial ones, this approach becomes quickly infeasible. Typically, only few of all possible features are present in a combinatorial fingerprint and often the total number of potential features might be impractical to enumerate. Therefore, by subtracting the constant term log (q_i^A/q_i^B) from each likelihood, log R(v_i) features not present are assigned a weight of 0 so that the score becomes

$$S(v_i) = \log R(v_i) - \log \frac{q_i^A}{q_i^B} = v_i \left(\log \frac{p_i^A}{p_i^B} - \log \frac{q_i^A}{q_i^B} \right) \tag{5}$$

and

$$S(\mathbf{v}) = \sum_{i=1}^{n} v_i \left(\log \frac{p_i^A}{p_i^B} - \log \frac{q_i^A}{q_i^B} \right) \tag{6}$$

The final equation shows that the Bayesian score leads to a weighting scheme for each bit position i that is dependent on the relative frequency of occurrence of the feature in active and database compounds. This weighting scheme is equivalent to the R4 weight for substructural analysis [19, 22]. The score for each compound is thus calculated by adding up the log-odds weights for each bit set on in the fingerprint of a test compound, while bits set off are ignored making the function easy to calculate for sparsely populated high-dimensional combinatorial fingerprints.

The estimates of the frequencies p_i^A are usually only based on a small sample size. In this case, it is important to smooth the distributions by applying an m-estimate correction [28].

$$\hat{p}_i^A = \frac{p_i^A N_A + m p_i^B}{N_A + m} \tag{7}$$

The weights in Eq. (6.6) are directly derived from the estimated probability distributions of active and database compounds and reflect the differences in the distribution of each feature. In order to assess the significance of each feature and ultimately estimate the expected recall performance, the discriminatory power of the features is quantified by considering the divergence between the feature distributions of active and inactive compounds. For this purpose, the Kullback–Leibler divergence, an information-theoretic measure to assess the difference between

probability distributions [29, 30], is applied. Given the assumption of feature independence, we obtain $P(\mathbf{v}|A) = \prod_{i=1}^{n} P(v_i|A)$ and $P(\mathbf{v}|B) = \prod_{i=1}^{n} P(v_i|B)$. By elementary manipulation, the Kullback–Leibler divergence given as

$$D(P(\mathbf{v}|A)||P(\mathbf{v}|B)) = \sum_{\mathbf{k}=(0\ldots0)}^{(1\ldots1)} P(\mathbf{v}=k|A)\log\frac{P(\mathbf{v}=\mathbf{k}|A)}{p(\mathbf{v}=\mathbf{k}|B)} \quad (8)$$

can be expressed as

$$D(P(\mathbf{v}|A)||P(\mathbf{v}|B)) = \sum_{i=1}^{n}\left(p_i^A\log\frac{p_i^A}{p_i^B} + q_i^A\log\frac{q_i^A}{q_i^B}\right) \quad (9)$$

Thus, the Kullback–Leibler divergence directly corresponds to the expected value of the score of an active compound by

$$\begin{aligned} E[S(\mathbf{v})] &= \sum_{i=1}^{n} p_i^A\left(\log\frac{p_i^A}{p_i^B} - \log\frac{q_i^A}{q_i^B}\right) \\ &= D(P(\mathbf{v}|A)||P(\mathbf{v}|B)) - \sum_{i=1}^{n}\log\frac{q_i^A}{q_i^B} \end{aligned} \quad (10)$$

If the screening database contains a small – yet unknown – number of compounds having activity similar to the reference molecules, the Kullback–Leibler divergence correlates with the percentage of active compounds among database compounds producing best scores according to the weighting scheme.

To utilize this correlation for the prediction of compound recall, it is required to quantitatively relate the Kulback–Leibler divergence to the expected recall performance. This can be achieved by generating a regression model based on a variety of different activity classes. Reference compounds from these classes are used in two ways. First, benchmark trials are performed where a portion of the compounds is "hidden" in a screening database and fingerprint search calculations are carried out to determine the recall rates for these compounds. These calculations can be carried out with methods of choice, for example, *k*-NN calculations in combination with the Tanimoto coefficient or Bayesian weights, as described above. In addition, the Kullback–Leibler divergence of the reference compounds is determined. Thus, for benchmark trials, two values are obtained: (1) the recall (for a certain percentage of the screening database) and (2) the Kullback–Leibler divergence of the reference set used for this trial. For each activity class, the benchmark trials should be repeated several times using randomized subsets as reference and as active database compounds. This yields a number of data points consisting of the Kullback–Leibler divergence and actual recall rates. For example, for 40 activity classes and 100 randomized trials per class, one would obtain 4,000 data points. Then a curve is fitted to these

data points and utilized to predict the recall performance of a new activity class only based on the Kullback–Leibler divergence retrieved from the bit distributions of the fingerprints of this class. A linear regression model of the logarithm of the Kullback–Leibler divergence and the recall of active compounds is found to yield high correlation and meaningful recall rate predictions for a variety of test classes, as discussed in the application section. In the next section, step-by-step instructions are provided how to generate linear regression models relating the Kullback–Leibler divergence to the (expected) recall performance.

2.2. Practical Prediction of Compound Recall

For a screening database, a fingerprint, and a set of compound activity classes, the steps required in order to generate a regression curve relating the Kullback–Leibler divergence to recall are the following:

1. Preprocessing: Fingerprint calculation and bit frequency determination.

   Initially, the fingerprints for all compounds in the database and the activity classes are calculated. Then the probabilities p_i^B for each bit position i are estimated from the relative frequencies.
2. Determination of the data points for the regression curve:

   For each activity class, a number of randomized similarity search trials are performed and the ratio of active compounds retrieved among a certain fraction of top-ranked database is determined. For practical applications, one should consider on the order of 100 trials per activity class where randomized subsets of about ten compounds are used as reference molecules and the remaining compounds are added to the database as potential hits.
   (a) For each trial, estimate the probabilities p_i^A for each bit position from the ten reference compounds using the relative frequencies and correct these values using the m-estimate according to Eq. (6.7). A value of 1 for m yields meaningful results in practice.
   (b) Determine the Kullback–Leibler divergence according to Eq. (6.9).
   (c) Perform a similarity search using the ten reference compounds and determine the recall of the active compounds in, for example, the top 1% of the database (i.e., determine the ratio of potential database hits within the top-ranked 1%).
3. Derive the linear regression curve:

   Step 2 yields about 100 data points consisting of a Kullback–Leibler divergence and a recall rate per activity class. For the

pooled data from all activity classes, determine a linear regression model by relating the logarithm of the Kullback–Leibler divergence to the recall rate.

4. Predicting compound recall for an activity class not included in the regression analysis:
   For a set of active compounds, determine the relative frequency distributions of each feature and apply the m-estimate correction according to Eq. (6.7). Determine the Kullback–Leibler divergence according to Eq. (6.9). Use the regression curve derived in **step 3** to predict the recall performance for a given fingerprint. It should be noted that expected recall rates are predicted under the assumption that active compounds similar to reference molecules are contained in the database. No information is available if the database indeed contains compounds having the desired activity. The estimated recall rate only predicts the percentage of active compounds that might be expected to be retrieved for the specific activity class and fingerprint. However, if a fingerprint has low predicted recall, it would not be able to retrieve such molecules, if available. Consequently, a different fingerprint should be tested.

2.3. Application Example

In order to test the applicability of the approach, reference calculations were carried out using two different fingerprints implemented in the Molecular Operating Environment (MOE) software [31]. The first one is a three-point pharmacophore-type fingerprint termed TGT (Typed Graph Triangle). It distinguishes four different atom types and six different distance ranges, comprising a total of 1,704 bits. The other is the structural fragment fingerprint MACCS encoding 166 predefined structural features [32]. These two fingerprints were tested with three different search strategies: the centroid approach [16], 5-NN calculations [16], and the Bayesian approach presented above, which corresponds to the R4 bit weighting scheme from substructural analysis [19, 22]. A centroid fingerprint of an activity class is calculated by averaging the individual fingerprints of all reference compounds. In 5-NN search calculations, similarity scores are averaged for each test compound over the five most similar reference molecules. Thus, centroid and 5-NN search strategies operate at different levels, either by fingerprint modification (centroid) or by similarity scoring (5-NN). Moreover, R4 weighting differs from both strategies in that it is based on relative bit frequencies of active and inactive compounds.

Calculations were carried out on an in-house generated subset of the ZINC [33] database that contained ~1.44 million compounds having unique 2D graphs. Thirty-nine different compound activity classes were used to study the relationship between Kullback–Leibler divergence and compound recall rates

consisting of 14–159 compounds [25]. For each class, 100 trials using randomized subsets of ten reference compounds were carried out, while the remaining compounds were used to determine the recall performance. All ZINC compounds were considered decoys. Recovery rates were calculated for the top 100 compounds of the ZINC database and related to the Kullback–Leibler divergence of the respective reference sets.

For the six similarity search methods (two fingerprints combined with three search strategies), linear regression models were created according to the protocol described above. The results are shown in Fig. 1 showing the regression model and the average recall performance of the 39 activity classes used for calibration of the curve. Recovery rates varied significantly among the 39 activity classes. Given its design, the Bayesian approach is most suitable for establishing a correlation between compound recall and Kullback–Leibler divergence. However, a significant correlation for the alternative search strategies was also observed. The chosen search strategy in part influenced the degree of the observed correlation. Linear regression models with correlations varying between 0.51 and 0.75 were obtained and average prediction errors ranged from ~16% to ~18%.

In order to test the ability of the models to predict the recovery rates of other activity classes, seven additional classes were used consisting of 16–71 compounds [25]. Details for these classes are given in Table 1. In order to obtain statistically sound results, each of these seven classes was subjected to 100 trials by splitting the class into randomized subsets of ten reference compounds, for which the Kullback–Leibler divergence was determined. For the remaining compounds, recovery rates were determined. The observed recovery rates were then compared to recovery rates predicted by the regression models, as shown in Fig. 2. Table 2 reports the predicted and observed recovery rates. As shown in Fig. 2, predicted and observed recovery rates varied from method to method. On average, the predicted rates deviated from the observed recall by 13–21%. These values were comparable to the observed average error of the training classes. Overall, recovery rates were well predicted for classes producing low as well as high recovery rates. Although the variation in performance varied in part significantly depending on the activity class and method, meaningful predictions were obtained in almost all cases (see also Note 3). Recall performance for each activity class much depended on the reference molecules chosen for each trial. Significant variations in recall performance between individual trials were observed, frequently ranging from ~10% to ~20% [34]. Prediction errors fell into the same range, thus making predictions of recall performance possible within the margin of deviations between individual trials.

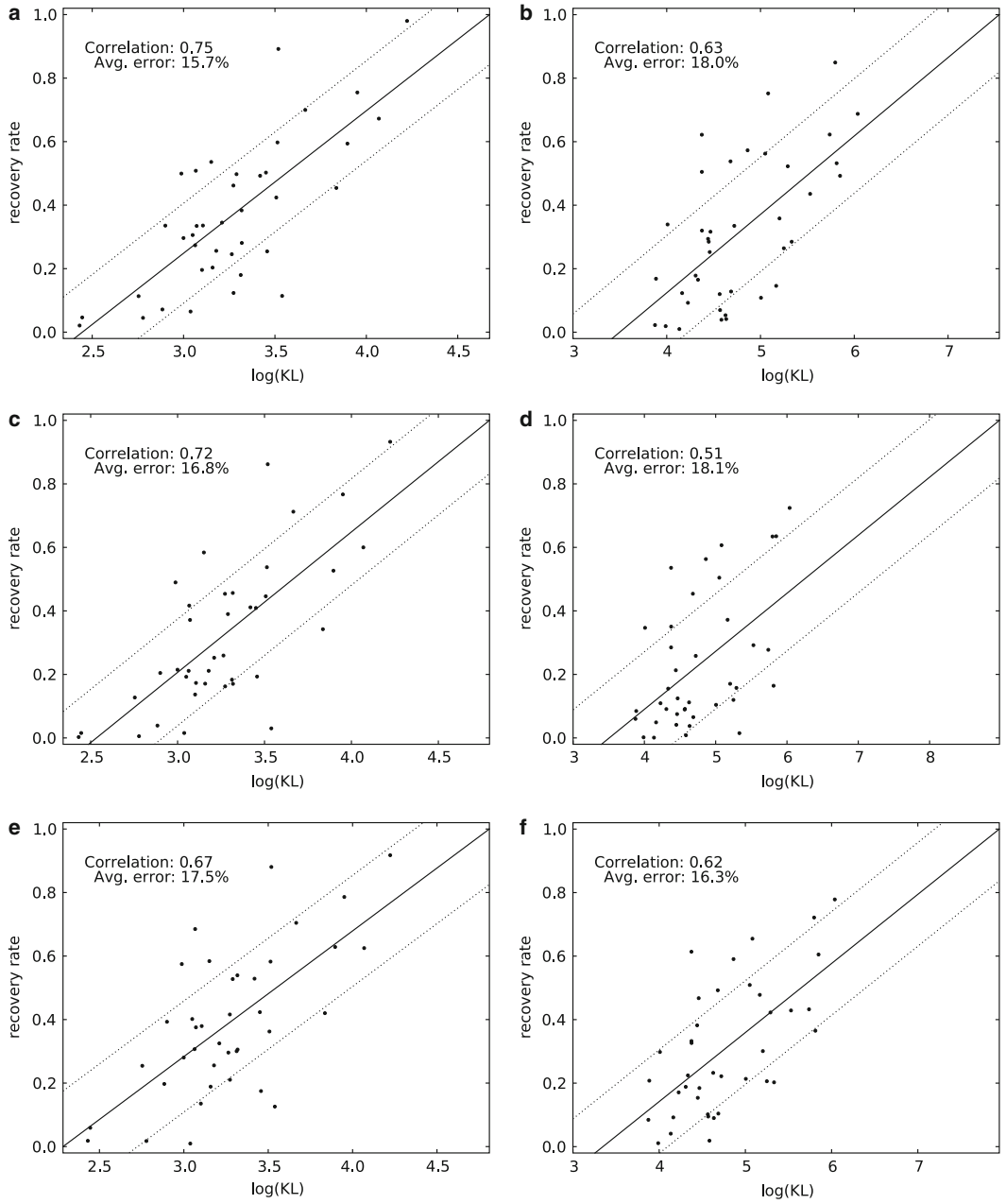


Fig. 1. Linear regression analysis of training data. For 39 different compound activity classes, average recovery rates from 100 individual trials were calculated and plotted against the logarithm of the corresponding Kullback–Leibler divergence for the MACCS (**a**, **c**, **e**) and TGT (**b**, **d**, **f**) fingerprint using the R4 (**a**, **b**), centroid (**c**, **d**), and 5-NN (**e**, **f**) search strategies. The *solid lines* represent the linear regression curves and the *dotted lines* the average error margins.

Table 1
Compound activity classes used to predict recovery rates. In order to assess intraclass structural diversity, average values of the Tanimoto coefficient (avTc) were calculated for pairwise comparison of compounds using the set of 166 publicly available MACCS structural keys

Class designation	Biological activity	No. of compounds	avTc
5HT	5-HT serotonin receptor ligands	71	0.67
ADR	β-Receptor antiadrenergics	16	0.74
ARI	Aldose reductase inhibitors	24	0.47
BEN	Benzodiazepine receptor ligands	59	0.69
DD1	Dopamine D1 agonists	30	0.57
KAP	κ Agonists	25	0.57
XAN	Xanthine oxidase inhibitors	35	0.56

3. Notes

1. The methodology described herein is in principle applicable to any fingerprint design, with two exceptions: (a) value range encoding fingerprints [9, 11], where the value range of a single chemical descriptor is encoded over a segment of bits and (b) hashed fingerprints such as the Daylight fingerprints [8], where molecular properties or connectivity patterns are mapped to overlapping bit segments. In both cases, individual subsets of bit settings depend on each other.
2. Compared to Bayesian modeling of continuous value distributions [26, 27], the fingerprint bit settings provide a significant advantage. The binary distribution capturing bit settings is discrete and does not require to making the critical assumption that feature values follow a normal distribution.
3. Similarity search performance generally depends on the nature of compound classes, chosen reference compounds, and molecular representations (e.g., fingerprints). Therefore, it is desirable to compare different fingerprints and their expected similarity search performance. A recent study evaluated different fingerprints by Kullback–Leibler divergence analysis and demonstrated the ability of the approach to select preferred representations for individual compound classes [35].

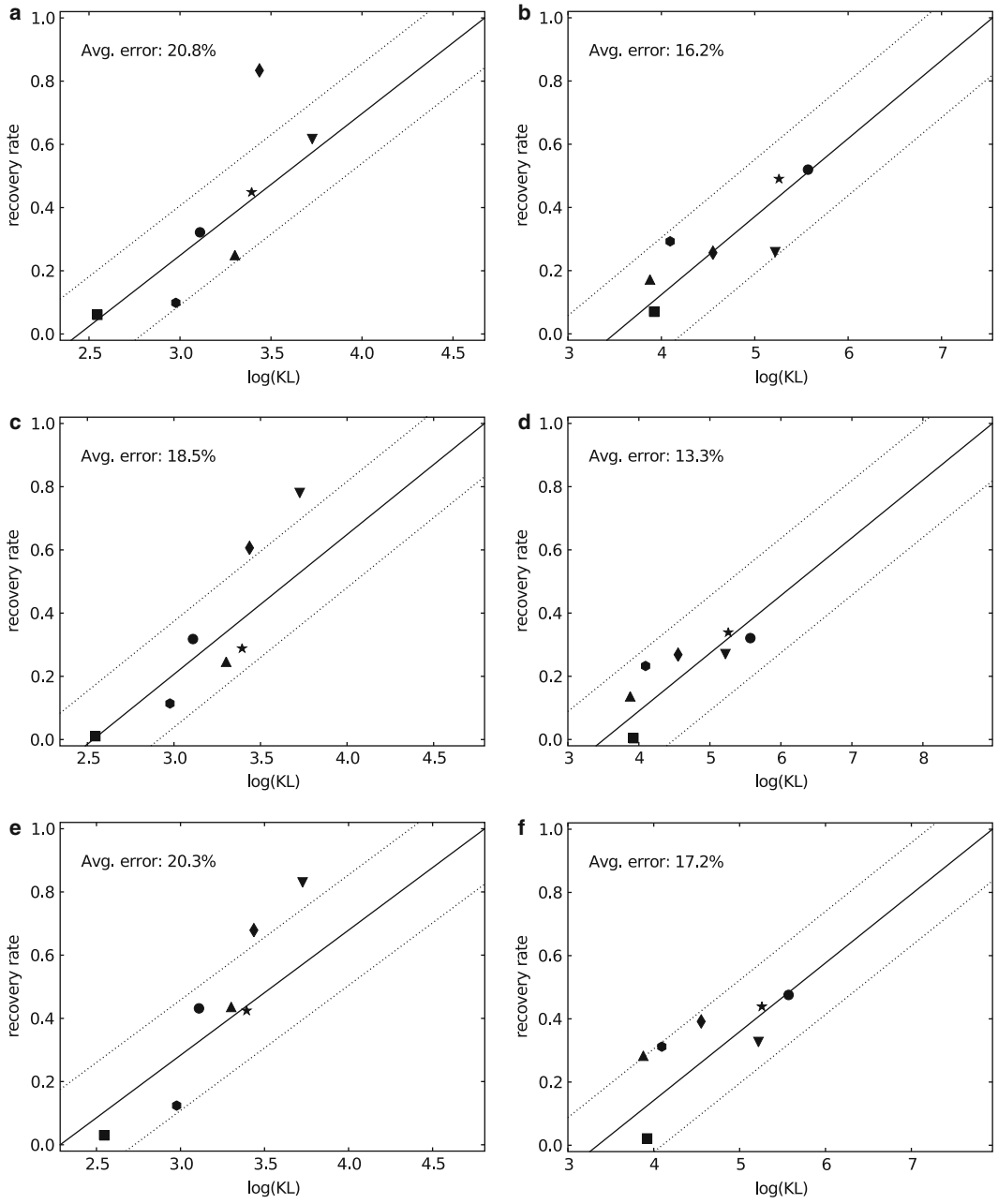


Fig. 2. Prediction of average recovery rates. For seven activity classes, recovery rates were calculated and compared to rates predicted by the regression model. Recovery rates are reported as averages over 100 individual trials and database selection sets of 100 compounds. The presentation is according to Fig. 1. The seven classes are represented as follows: *triangle*, 5HT; *inverse triangle*, ADR; *square*, ARI; *diamond*, BEN; *star*, DD1; *hexagon*, KAP; *circle*, XAN.

Table 2
Predicted and observed recovery rates for the seven test classes. Recovery rates (RR) were predicted from Kullback–Leibler divergence and averaged over 100 individual trials for selection sets for 100 database compounds. Observed recovery rates were also averaged over 100 trials. Predicted recovery rates are given in parentheses. Furthermore, the average prediction error over 100 trials is reported

Class designation	R4		Centroid		5-NN	
	Avg. RR (predicted RR)	Avg. error	Avg. RR (predicted RR)	Avg. error	Avg. RR (predicted RR)	Avg. error
MACCS						
5HT	25.0 (38.4)	13.7	24.6 (34.0)	10.3	43.6 (40.3)	15.0
ADR	61.7 (57.5)	14.3	78.0 (52.8)	26.5	83.0 (57.1)	26.4
ARI	6.1 (4.5)	9.1	1.1 (0.5)	6.8	3.0 (10.4)	9.5
BEN	83.4 (44.4)	38.9	60.6 (39.9)	23.6	67.9 (45.6)	24.7
DD1	44.9 (42.5)	11.4	28.9 (38.0)	11.5	42.5 (43.9)	10.3
KAP	9.9 (23.8)	15.9	11.4 (19.6)	11.8	12.4 (27.4)	16.3
XAN	32.2 (29.8)	10.4	31.8 (25.5)	9.6	43.2 (32.7)	12.7
TGT						
5HT	17.2 (9.4)	11.4	13.7 (6.8)	10.7	28.3 (11.5)	17.0
ADR	25.8 (42.5)	20.1	27.0 (31.3)	14.0	32.7 (40.7)	16.4
ARI	7.0 (10.5)	6.6	0.4 (7.6)	7.2	2.1 (12.5)	10.5
BEN	25.7 (26.0)	10.4	26.8 (19.1)	9.7	39.2 (26.2)	15.4
DD1	49.0 (43.5)	10.8	33.9 (32.0)	8.8	43.9 (41.6)	12.8
KAP	29.3 (14.7)	19.6	23.3 (10.7)	14.6	31.2 (16.2)	17.1
XAN	52.0 (51.2)	9.4	32.1 (37.7)	10.3	47.6 (48.3)	11.3

References

1. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
2. Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nature Rev. Drug Discov.* **1**, 882–894.
3. Willett, P. (2005) Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **48**, 4183–4199.
4. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053.
5. Barnard, J. M. and Downs, G. M. (1997) Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **37**, 141–142.
6. Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280.
7. *MACCS Structural Keys.* Symyx Technologies, Inc., Sunnyvale, CA, http://www.symyx.com (accessed Sep 1, 2009).
8. James, C. A, Weininger, D. *Daylight Theory Manual*, Vers. 4.9, Daylight Chemical

Information Systems Inc., Aliso Viejo, CA, http://www.daylight.com/dayhtml/doc/theory (accessed Sep 1, 2009).
9. Xue, L., Godden, J. W., Stahura, F. L., and Bajorath, J. (2003) Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **43**, 1151–1157.
10. Bender, A, Mussa, Y, Glen, R. C., and Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **44**, 1708–1718.
11. Eckert, H. and Bajorath, J. (2006) Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J. Chem. Inf. Model.* **46**, 2515–2526.
12. Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., and Labaudiniere, R. F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview over the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **42**, 3251–3264.
13. Bradley, E. K., Beroza, P., Penzotti, J. E., Grootenhuis, P. D. J., Spellmeyer, D. C., and Miller, J. L. (2000) A rapid computational method for lead evolution: description and application to α_1-adrenergic antagonists. *J. Med. Chem.* **43**, 2770–2774.
14. Maggiora, G. M., and Johnson, M. A. (1990) *Concepts and Applications of Molecular Similarity*. Wiley: New York, NY, pp 99–117.
15. Hert, J., Willet, P., and Wilton, D. J. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **44**, 1177–1185.
16. Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003) Similarity metrics for ligands reflecting the similarity of the target protein. *J. Chem. Inf. Comput. Sci.* **43**, 391–405.
17. Whittle, E., Gillet, V. J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: theoretical model. *J. Chem. Inf. Model.* **46**, 2193–2205.
18. Whittle, E., Gillet, V. J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: similarity searching and group fusion. *J. Chem. Inf. Model.* **46**, 2206–2219.
19. Hert, J., Willett, P, and Wilton, D. J. (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **46**, 462–470.
20. Lewis, D. D. (1998) Naïve (Bayes) at forty: the independence assumption in information retrieval. In *Lecture notes in computer science: Machine learning ECML-98*, Springer: Berlin, 4–15.
21. Zhang, H. (2004) The optimality of naïve Bayes. In *Proceedings of the seventeenth Florida artificial intelligence research society conference*. The AAAI Press: Menlo Park, CA, 562–567.
22. Ormerod, A., Willett, P., Bawden, D. (1989) Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Act. Relat.* **8**, 115–129.
23. Eckert, H. and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations, and novel approaches. *Drug Discov. Today* **12**, 225–233.
24. Sheridan, R. P. and Kearsley, S. K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* **7**, 903–911.
25. Vogt, M. and Bajorath, J. (2007) Introduction of a generally applicable method to estimate retrieval of active molecules for similarity searching using fingerprints. *ChemMedChem* **2**, 1311–1320.
26. Vogt, M., Godden, J. W., and Bajorath J. (2007) Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *J. Chem. Inf. Model.* **47**, 39–46.
27. Vogt, M. and Bajorath, J. (2007) Introduction of an information-theoretic method to predict recovery rates of active compounds for Bayesian in silico screening. *J. Chem. Inf. Model.* **47**, 337–341.
28. Berthold, M. and Hand, D. J. (2007) *Intelligent Data Analysis: An Introduction*. Springer: Berlin, Heidelberg, Germany, pp 245–246.
29. Kullback, S. (1997) *Information Theory and Statistics*. Dover Publications: Mineola, MN, pp. 1–11.
30. Cover, T. M., Thomas, J. A. (1991) *Elements of Information Theory*. Wiley-Interscience: New York, NY, pp. 224–238.
31. *Molecular Operating Environment (MOE)*, Vers. 2005.06, Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3, http://www.chemcomp.com (accessed Sep 1, 2009).
32. McGregor, M. and Pallai, P. (1997) Clustering of large databases of compounds: using the MDL "keys" as structural descriptors. *J. Chem. Inf. Model.* **37**, 443–448.
33. Irwin, J. J. and Shoichet, B. K. (2005) ZINC – A free database of commercially available

compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182.
34. Vogt, M. and Bajorath, J. (2008) Bayesian screening for active compounds in high-dimensional chemical spaces combining property descriptors and fingerprints. *Chem. Biol. Drug Design* **71**, 8–14.
35. Vogt, M., Nisius, B., and Bajorath, J. (2009) Predicting the similarity search performance of fingerprints and their combination with molecular property descriptors using probabilistic and information-theoretic modeling. *Stat. Anal. Data Mining* **2**, 123–134.

Chapter 7

Bayesian Methods in Virtual Screening and Chemical Biology

Andreas Bender

Abstract

The Naïve Bayesian Classifier, as well as related classification and regression approaches based on Bayes' theorem, has experienced increased attention in the cheminformatics world in recent years. In this contribution, we first review the mathematical framework on which Bayes' methods are built, and then continue to discuss implications of this framework as well as practical experience under which conditions Bayes' methods give the best performance in virtual screening settings. Finally, we present an overview of applications of Bayes' methods to both virtual screening and the chemical biology arena, where applications range from bridging phenotypic and mechanistic space of drug action to the prediction of ligand–target interactions.

Key words: Bayes Classifier, Virtual screening, Structure-activity relationships, Mode of action analysis, Target prediction, Adverse drug reactions

1. Introduction

Pharmaceutical companies, and life science research in general, generate amounts of life science data which is both enormous in size and genuinely heterogeneous (chemical, biological, text) in nature. While individual data points can be of tremendous value – and make the difference between a drug on the market and no drug on the market – most data generated can only be transformed into *information* and finally into *knowledge* when looking at multiple pieces of data in parallel. This is the field where data mining algorithms come into play, with two major aims whose relative importance differs on the particular setting: on the one hand, data mining tools are aiming to help interpret the data; on the other hand, those tools can also under certain conditions (such as sufficient coverage of input variables) be used to predict experimental quantities of novel matter.

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_7, © Springer Science+Business Media, LLC 2011

In this chapter, we focus on a particular set of data mining methods, namely, those based on "Bayes' Theorem." This theorem, published originally by eighteenth century statistician Thomas Bayes [1], can be used to model the probability distribution of an output variable, if conditional probabilities (in practical terms often relative frequencies) of the input variables for a set of classes are known. While this may seem trivial, this theorem is often counterintuitive to humans – such as in the case of analytical tests of rare diseases, where a 99% true positive rate in a clinical test, at a disease frequency of 1:100,000, means that for every person, where the detection of this disease takes place *a thousand* people will get false-positive test results. Hence, sticking to the formulation of Bayes' theorem might seem sometimes counterintuitive to humans, but its validity has been shown both in theory and in practical applications.

Deriving Bayes' theorem is not a difficult task, since we can write that the conditional probability of an event A, given event B, can be written as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

In the same way, the probability of event B, given event A, can be written as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

The joint probability, $P(A \cap B)$, is common to both equations; hence, after rearranging and equating the remainder we arrive at:

$$P(A|B)P(B) = P(B|A)P(A),$$

which can in turn be rearranged to give the commonly known form of Bayes' theorem, namely,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}.$$

In practice, most often the "event B" is the presence or absence of particular features of an item (such as a molecule); the predicted likelihood of "event A" is then a property of this item (such as the bioactivity of the molecule). Hence, using the Bayes' theorem, from the distribution of features in a dataset with different classes, the probability of a class membership, given certain features, of a novel item can be predicted.

Note that $P(A)$, called the "prior" in Bayes' nomenclature, makes an assumption about the likelihood of A, in the absence of any additional information. In practice, this entity is often not trivial to estimate; in many cases, information from a given

dataset can be used to make an estimate of the prior (e.g., if 5% of all items in a dataset show a particular property, the prior will also be set to this value). However, in some situations such as when, in virtual screening, ranking compounds relative to each other, the prior will be identical for every compound, and in practice in those situations a uniform prior ($P(A) = 1$) might be used for simplicity.

Given that in many practical cases the probability $P(B)$ may be zero for particular properties not encountered in the training set, this exception needs to be treated differently. Likewise, very small sample sizes can be misleading – for example, in cases where a feature is only present once in an active molecule, but never in inactive molecules, the likelihood according to the Naïve Bayes Classifier that a new molecule showing that feature is unity (100%). In order to remedy both of the above situations in practice, the Laplacian Correction is employed for the Naïve Bayes Classifier, which includes adding k "virtual" samples to the above formula with the assumption that in each of those virtual samples a particular feature will be present. Various forms of this Laplacian Correction term exist though, and for a detailed analysis the reader is referred to recent comprehensive studies [2, 3].

In practical settings, both regression and classification tasks can be solved using Bayes' approaches; however, classification tasks are much easier to solve in practice. This is relatively easy to do in cases, where also categorical input variables are present; if numerical inputs are present the so-called *binning* of input variables can be performed. Various different approaches exist here, with the most common ones being either fixed distances between bin borders (which is simpler to perform), or the aim to have equally populated bins in the final feature vectors (which transports the highest amount of information, from an information-theoretic point of view). In practice, although using equipopulated bins seems to give slightly improved performance in many situations [4], fixed bin borders are used however [3], and for further details the reader is referred to the studies cited here.

In classification tasks, often the Naïve Bayesian Classifier is employed, which gets its name from the fact that it only considers the frequency of individual features in the classification, but not their mutual dependence. This can have tremendous effect in practice, since the classifier assumes that every input variable increases the confidence in a particular category assignment of an item, while in reality all the variables are based on the same hidden factor and thus only marginal confidence in a class assignment can be given. However, a detailed study gave an interesting result on the performance of the Bayes Classifier under different degrees of correlation of the input variables: It was found that both *no correlation* as well as *high correlation* between input variables leads to good results of Naïve Bayes; however, intermediate

correlation caused frequent misclassification [5, 6]. In the context, we use the Bayes Classifier in here, namely, virtual screening, its satisfactory performance in case of intermediate correlation between features could be a reason for concern, since very frequently features used to type molecules, such as fingerprints, are often highly correlated (the same features are used multiple times to generate features, such as circular fingerprints, hence they contain to a good extent very similar – or correlated – information). In practice, though, this has not been found to be a problem, with Bayes' methods being very well-performing in comparative virtual screening studies (*see* also following section).

The task we apply to the Bayes Classifier in this chapter is mainly the *virtual screening* of ligand libraries [7]. This means the computational (in silico) selection of putative protein ligands, based on a set of known active compounds. There are multiple aims to this step; it can be used to find novel scaffolds active against the same target (and being a superior starting point for lead optimization efforts), to explore the SAR around a particular hit identified by screening additional, similar compounds; or to find compounds that are more active or that possess, for example, improved ADME/Tox properties, compared to the query structure. In every case, the two essential steps in the virtual screen are on the one hand the *description* of the molecule, and on the other hand the particular method used to compare the two (or more) descriptor vectors between the library molecules, and the one or more active query compounds [7]. While in this chapter we do not focus on molecular representations, it should be mentioned that the Bayes Classifier is suited for combination with most fingerprinting methods available, and that also very large feature sets ($\gg$ millions of features) can be handled efficiently by this method (such as in case of combination with e.g., circular fingerprints which encompass huge feature spaces).

In general, it can be said that while other machine learning methods perform sometimes superior to the Bayes Classifier [8], its true strength lies in the combination of very good performance, time-efficient learning, and its ability to handle multiple categories ($>$ thousands) efficiently. The latter is particularly useful in "reverse" virtual screens, where for a given ligand its potential protein interaction partners should be identified and where those possible targets are at least hundreds, but given appropriate databases even thousands of categories (*see* next section).

The time-efficiency of the Bayes Classifier stems from its character as a "single-shot" learner – in effect, all it does is calculating and storing feature frequencies for each individual feature, applying (if necessary) a correction factor, such as the Laplacian Correction, and the model generation process is finished. This is in sharp contrast for iterative learning methods such as Support Vector Machines and Neural Networks, which often go through

hundreds of iterations to establish a final model. In addition, those other methods often involve random elements (such as the initialization of the node weights in case of neural networks), making the result often not reproducible by identifying multiple local minima in the output function – in case of the Bayes Classifier, feature frequencies are given for a particular dataset, and depend neither on initialization of the method (which is not necessary), nor on the order of data points used for training the model (since the order of data does not influence conditional probabilities of features).

One thing that needs to be kept in mind when using Bayes classification techniques is that the output variable is trained in a binary manner, with the (numerical) output values not giving a numerical value of a property (such as bioactivity), but only the *likelihood* of a molecule to show this character (e.g., being active above a particular, by and large arbitrary, threshold). This is understandable from the way a Bayes Classifier is trained, where only the categorical input labels are provided for training, and the numerical output stemming only from the frequency of features in each of the classes. Hence, the potency or efficacy of a compound and its likelihood of being active (at any potency) are, strictly speaking, two completely independent phenomena. That being said, in practice of course often the presence of multiple ligand features binding to a protein is correlated with binding affinity to some extent, but this is certainly not true in each individual case.

While the Naïve Bayesian Classifier only considers the conditional probability of individual variables, given a particular function output, the mutual dependency of input variables is neglected completely. However, this is only a valid approximation in very rare cases, and virtually never for the tasks of virtual screening performed here: ligand chemistry shows highly repetitive (and "correlated") chemistry [9] on the one hand, and molecular fingerprints are often calculating from overlapping fragments/paths on the other hand [7]. Hence, more recently also Bayesian Inference Networks, which consider the conditional probability of feature combinations explicitly, have been introduced into the realm of virtual screening with considerable success [10], and we discuss particular applications in the following section.

The application of the Bayes Classifier – and the whole, sometimes emotional debate of "Bayesian vs. Frequentist methods," has also had its influence on discussions on how industry should handle data originating from large HTS campaigns [11] (as also in other areas, such as how to treat results from clinical studies with the appropriate statistical tools [12, 13]). Given the decreasing number of new chemical entities (NCEs) approved by the FDA in recent years, the nature of pharmaceutical drug discovery where over time more and more data becomes available and at each stage "the best possible choice given the data" needs to be made, the

importance of this question becomes obvious. In addition to the above review [11], the interested reader is referred to a recent review dealing with Bayesian modeling in virtual high-throughput screening [14], which lists a variety of recent practical applications of this technique in the drug discovery field.

2. Applications of Bayes' Methods in Virtual Screening

While Bayes' theorem is already more than two centuries old [1], its applications in the area of virtual screening only appeared in the last decades. One of the first works in the area explicitly mentions the application of the Bayes Classifier for compound classification within the "Binary QSAR" framework [15]. In this original contribution, the input variables into the classifier were considered as real-valued variables, which were firstly binned (assigned to fixed property ranges). Subsequently, a constant factor was added (to avoid a zero input density), and Gaussian smoothing was applied to estimate property densities separately for the "active" and the "inactive" dataset. In this work, the method was tested on a set of 1,947 molecules with molar refractivity data which were encoded by a set of four connectivity indices. While no comparison to other methods was performed, classification accuracy in the region of up to above 90% point to good performance of the method. Even more interestingly, in order to judge the behavior of Binary QSAR in difficult situations, it was found that for uniformly distributed errors as well as random wrong assignment of class membership in the training data still excellent performance of the method was obtained. This hints at good performance of the method also when significant noise in the data is present (such as in HTS data which was confirmed later, *see* below). However, in case of a very small number of active data points the performance of the method significantly suffered, underlining the importance of having a sufficiently large number of training data at hand to obtain meaningful feature density distributions for the generated model. (This result was confirmed on an additional dataset recently, mentioning group fusion as a method superior to Naïve Bayes in cases, where only very few active datapoints are available [16]. "Very few," in the personal opinion of the author, probably refers to about a dozen or so active compounds in practice – above this number of datapoints, as a rule of thumb, the Bayes Classifier should give good performance in the typical virtual screening situation as shown in large-scale retrospective virtual screening studies [17].) In addition to the original study, the method was applied to the detection of Estrogen receptor ligands [18] as well as natural products [19] from large compound collections, and classification accuracies of better than 80% and 94% were obtained in the above

studies, respectively. In addition, the generation of focused libraries for cyclic GMP phosphodiesterase type V inhibitors and for acyl-CoA:cholesterol O-acyltransferase inhibitors was performed with the same method [20].

While in the approach above, binning was performed on numerical input descriptors, more recently also real-valued input was modeled as univariate and multivariate distributions, respectively [8, 21]. While the univariate treatment of input descriptors does not consider mutual dependence between variables (in line with the above mentioned Naïve Bayes Classifier), in multivariate input approaches feature densities are obtained which do capture the mutual information of two (or more) input variables, most commonly approximated as a multidimensional Gaussian function [21].

Also a free implementation of circular fingerprints in combination with the Naïve Bayes Classifier has been published and benchmarked, termed "Molprint 2D" [17, 22]. Multiple studies hint at the ability of circular fingerprints such as PipelinePilot ECFP fingerprints, or the just mentioned Molprint 2D fingerprints, to capture molecular features associated with bioactivity to a degree superior to many other molecule classification methods [23, 24]. A benchmark comparison of different fingerprint types in combination with machine learning methods [24] is presented in Fig. 1, where the Bayes Classifier (first row) is among the virtual screening methods giving the highest retrieval on this particular dataset, comprising 11 bioactivity classes.

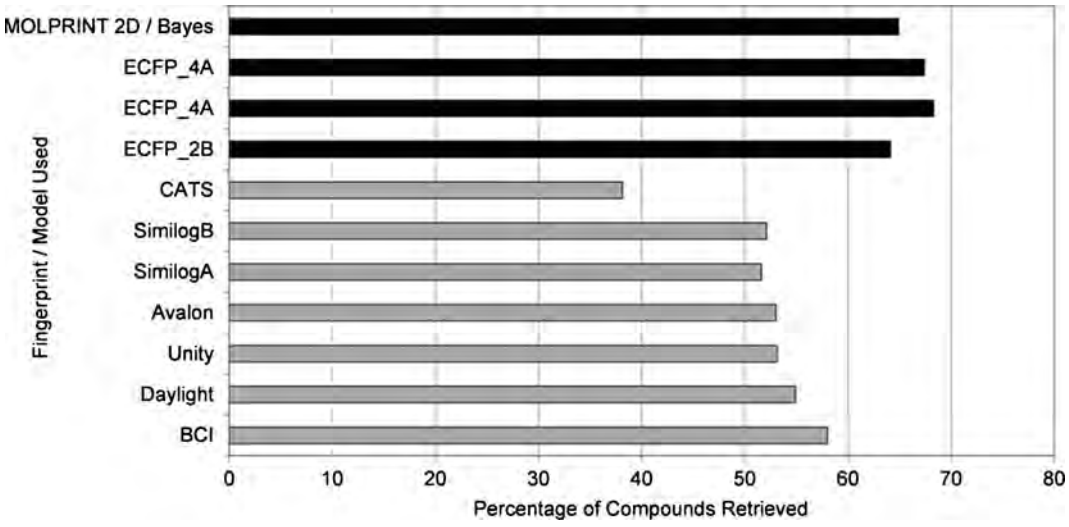


Fig. 1. Benchmark comparison of different fingerprint types in combination with machine learning methods. The Bayes Classifier (*first row*) is among the virtual screening methods giving highest retrieval on this particular dataset, comprising 11 bioactivity classes.

Given the ability of the Naïve Bayes Classifier to handle very large feature spaces (and the space of a circular fingerprint with, for example, 30 different atom types, arranged in a central atom with a radius of 2 into every direction, might extend to the range of about $20 \times (20^2) \times (20^4)$ or about 1.3×10^9 possible features), its performance in combination with the information-rich circular fingerprints can be expected to be outstanding; this has indeed shown to be the case in retrospective virtual screening studies. In one case [17], both circular fingerprints trained on single molecules and Binary Kernel discrimination trained on multiple molecules were outperformed by the Bayes Classifier on 11 sets of active compounds from the MDDR database. The "Molprint 2D" approach mentioned above has also been extended to molecular surfaces, choosing instead of individual atoms GRID-generated interaction potentials with a variety of probes as the descriptor features [25]. However, while giving results often orthogonal to established (and 2D) methods, the calculation of the "Molprint 3D" descriptors is a rather time-intensive step, requiring tens of milliseconds on today's machines per molecule.

In practical terms, the application of the Bayes Classifier in virtual screening has been greatly helped by its implementation (in combination with circular fingerprints) in Accelrys' PipelinePilot software, thus being a standard tool for activity model generation throughout industry.

Given the assumption the Naïve Bayes Classifier makes, namely, the mutual independence of features used for classification, the impact of feature selection methods on its performance in virtual screening [17] is not too surprising. In a detailed manner, recently the performance of the Bayes Classifier in combination with different feature selection methods, namely, information gain, mutual information, chi-squared test, odds ratio, and the GSS test (a simplified version of the chi-squared test) has been compared [26]. It was found that sensitivity could be increased significantly when removing up to 96% of the 139,351 features present in the dataset used in the "2001 KDD cup." Quite to the contrary, SVMs were able to perform rather well on the full-dimensional feature set, with performance dropping of sharply when fewer and fewer features are selected. This dataset comprised 1,909 thrombin inhibitors of which only 42 are active, hence this dataset could be seen as typical of HTS datasets, where also only a small number of positive data points are expected (although the total number of data points will usually be considerably higher).

Interestingly, the Naïve Bayes Classifier has been most often employed in combination with information-gain based feature selection in the area of virtual screening and compound classification [17, 27, 28], and as shown above this combination indeed

shows superior performance to no feature selection at all. However, the above study [26] points to the fact that odds ratio-based feature selection might give superior performance with respect to sensitivity than information gain-based feature selection; a finding that still should be validated on additional datasets with practically relevant characteristics in the future.

While virtual screening is most commonly based on molecular fingerprints that encode the chemical structure of molecules, also the differences of other properties such as physicochemical properties between active and inactive compounds can be analyzed using Bayesian methods. In this work [29], it was found that molecules with eight hydrogen bond acceptors, five hydrogen bond donors, one molecular weight between 550 and 600, and six rotatable bonds were up to a factor of 2 (in case of six rotatable bonds) more likely to be active in one of the analyzed assays, compared to the average overall hit rate. While some of the factors – such as larger molecules, being more lipophilic on average, have higher hit rates – have been discussed before, it is also evident from this work that molecular properties that confer bioactivity in assays does not necessarily resemble molecular properties required to bring a successful, orally available drug on the market.

As for novel developments in Bayes' methods applied to virtual screening, mainly two topics shall be touched upon here. On the one hand, this is "Bayesian Model Averaging," a technique introduced in the machine learning world [30] recently and which now also found its way into the area of virtual compound screening [31]. In this particular study, Bayesian model averaging was compared to Support Vector Machines and Artificial Neural Networks for the prediction of protein pyruvate kinase activity using DRAGON descriptors. Bayesian models were averaged over an ensemble of compound classification trees, and it was found that the resulting models were interpretable, in addition to showing the performance "at least as good if not better [than] SVM and Neural Networks" [31].

On the other hand, a recent novel development of Bayes' methods applied to virtual screening is the use of not only the Naïve Bayes Classifier to compound classification, but also considering the mutual dependence of features, resulting in a Bayesian Inference Network. [10] In this study, retrieval of active compounds from each of 12 activity datasets derived from the MDDR database was increased by 2–4% in absolute terms (or about 8–10% in relative terms) using a variety of circular count fingerprint-based methods, compared to the benchmark Tanimoto coefficient. This was unfortunately not true for the retrieval of different frameworks from the dataset; however, this result still hints at the importance of considering mutual dependencies of features in the virtual screening process. This can easily be understood since molecular features commonly are not responsible for

bioactivity by themselves, but they rather occur in *sets* of interaction features responsible for binding, and those sets differ, for example, in different chemical series, with different binding modes. Hence, the above result is not only interesting in the practical sense, but also appealing regarding our understanding of ligand-protein interactions.

3. HTS Follow-Up and Activity Modeling Using the Bayes Classifier

3.1. Follow-Up of Experimental HTS Data

High-throughput screening, the "brute force" approach of testing in the order of one million compounds for activity against one target, has become fashionable in the last two to three decades with increasing target-focused drug discovery activities. However, it is rare that a first HTS will give a suitable lead compound directly; rather, there will be (hopefully) multiple active hit series being discovered, which can be used to investigate the SAR surface around those series further, and to pick additional compounds either from an in-house database or an external supplier.

In order to mimic this process, in 2005 the Screening Laboratory at McMaster University announced the "High-Throughput Screening Docking Competition," where interested parties were encouraged to prioritize compounds from a 50,000 compound test set, based on screening results from a training set of the same size screened for *Escherichia coli* dihydrofolate reductase inhibition [32]. It should be mentioned at this stage that the test set was significantly dissimilar in chemistry from the training set, making the application of some methods difficult and giving possibly different conclusions than in easier settings. However, extrapolating from one area of active chemistry to another, by using suitable molecular descriptors as well as machine learning methods, is certainly one of the aims of generating activity models, and so the study of how different tools behave in this "extreme" situation was certainly worth investigating.

Two Bayesian modeling methods entered the competition, one based on ECFP fingerprints combined with a Laplacian-modified Bayes Classifier [33] and the other based on Molprint 2D fingerprints in combination with a different Laplacian Correction term [34]. It was found that using both kinds of circular fingerprints results in prioritizing the test set approached the performance of computationally much more expensive 3D methods [33]. In addition, it could be shown that "averaging" the chemistry between training and test set [34] by scrambling both data sets improved prioritization of the (reshuffled) test set. This corroborates the finding in the beginning of this section that while Bayesian Classifiers are able to handle uniform noise rather well, they often show a significant drop in performance if either a very small

number of active compounds is present (in case of virtual screens that means if less than at least about a dozen compounds are given), or if systematic errors or insufficient coverage of the test set space are given in the training set.

The question of how a machine learning method handles noise is crucial in many areas, but not often as crucial as in case of the analysis of HTS data. This type of data is notoriously noisy, not only owing to the large number of data points that needs to be obtained at high speed and low cost at the same time, but also due to the underlying biology that is inherently noisy. In a recent study [8], the Laplacian-modified Naïve Bayes Classifier was compared in its performance on four sample HTS datasets to recursive partitioning as well as support vector machines, with respect to increasing levels of noise on classification performance when false positives and false negatives were added to the data. Overall, it was found that all three methods tolerated increasing levels of false positives remarkably well, even in cases where the ratio of misclassified compounds to true active compounds was as much as 5:1 in the training set. For false negatives, tolerance of noise up to a ratio of 1:1 was achieved. While SVMs outperformed the two other methods in classification performance slightly, training a support vector machine (and choosing its parameters!) on a dataset of this size is a time-consuming exercise. Interestingly, upon adding noise in a 1:1 ratio to the original inactive screening data, in case of the Bayesian Classifier on one dataset classification performance significantly *improved*, a phenomenon that is not unheard of in nonlinear systems [8]. It should be kept in mind, however, that noise in the above example is stochastic noise; hence, systematic deviations in the input data will still cause problems in compound classifications based on the resulting model. In a follow-up study, also screening in "pools" (multiple compounds per well) under noise levels between 81% and 91% was analyzed, and the resulting model still yielded between 2.6-fold and 4.5-fold enrichment of the true actives from the same dataset [35]. Hence, not only compound prioritization on future datasets can be performed with the Bayes Classifier, also hit deconvolution from pooled HTS screens is an application area, where it shows satisfying performance.

Also to another important target class, namely, kinases, Bayesian methods were recently applied with an additional twist: If one considers an orphan target, for which no inhibitors are known, can one use the so-called *chemogenomics* methods and train a ligand-based model on other inhibitors active against members of this family, and still provide significant enrichment of inhibitors for the kinase of interest? For the family of tyrosine kinases this could indeed be shown, where using circular fingerprints and the Bayesian Classifier as implemented in PipelinePilot [33] extrapolation of activity predictions to related kinases were possible to perform.

Likewise, two important properties of the Bayesian Classifier were noted in this work: Firstly, that proper coverage of the "inactive" space is beneficial to classification results of the Bayes Classifier. This can probably be understood in a way that, in case of insufficient coverage of inactive space, if a new, unknown molecule is to be classified, for too many of the features present, the Bayes Classifier cannot draw any conclusion from the features provided and thus stays "undecided" about the new item encountered. Secondly, it was found that the Bayes Classifier is able to merge features from multiple molecules and thus to perform "scaffold hopping" [36] in virtual screening settings, where novel bioactive structural cores are identified which are also active against the target of interest. This finding is in agreement with other work both in the virtual screening area [37], but also in other areas such as the analysis of fragment-based screening data using circular fingerprints and the Naïve Bayes Classifier [38]. In the latter case, the Bayesian method was able to draw conclusions about the bioactivity of more than 1 million compounds from the HTS deck, based on a fragment-based screen of 8,800 fragments. Hence, the combination of circular fingerprints (which cover only patches of the molecule) and the Naïve Bayesian Classifier (which is able to recombine features in novel ways) seems to be a good combination to draw, to some reasonable extent, conclusions about chemical space previously not encountered in the training set.

3.2. Ligand-Based Bioactivity Modeling and Understanding Bioactivity Space

While the above methods of analyzing high-throughput screening data are based on the Naïve Bayesian Classifier in combination with individual, discrete features, Bayesian methods have also been applied to more "conventional" QSAR modeling, where in a relatively small dataset (up to hundreds of data points) a relationship between structure and activity of a compound is investigated (with, ideally, the perspective of further optimization of the compound). One major problem in this type of work is the "curse of dimensionality," where a large number of possible descriptors (thousands) can be generated for a molecule. However, only part of them will be related to the output variable in a meaningful way, but one does not know a priori which ones. Hence, the input vector into a QSAR model will become very large in many cases. In particular, in combination with artificial neural networks (ANNs) this can become problematic, since this type of modeling method possesses a large number of degrees of freedom (variables), leading frequently to overfitting, as well as the algorithms becoming trapped in local minima. In order to remedy this problem, a decade ago *Bayesian regularized* artificial neural networks (BRANNs) were introduced in the development of QSAR models [39]. These models punish larger models automatically, hence aiming to provide a model that at the same time optimizes

performance and generalization ability (which is generally inversely related to the size of the model). In the study cited here, the application of BRANNs to modeling compounds active on benzodiazepine and muscarinic receptors has been performed, using a set of three topological indices as input descriptors. It was found that, indeed, BRANNs remedy many of the problems associated with ANNs: In the 30 runs performed for model generation very similar models have been obtained (which is rarely the case for "standard" ANNs due to random initialization of the node weights); in addition, significantly better model statistics were obtained, compared to PLS models (such as for the large benzodiazepine set of 245 compounds a standard error of prediction of 0.14 (vs. 0.21 for PLS), and a Q2 of 0.69 (vs. 0.28 for PLS)). Hence, also in the quantitative structure-activity modeling area Bayesian methods have their justification and future, prospective applications of this method should be interesting to hear about.

Understanding bioactivity space around a particular target is especially crucial in case of large target classes, such as GPCRs, but also, in case for example of cancer, the group of kinases with an estimated size of 518 target class members [40]. For a set of kinase inhibitors, the group at Lilly analyzed the fragment contribution, generated according to their "Dicer" method, to the bioactivity against a set of 36 kinases [41]. Their findings are relevant in multiple aspects: Firstly, by inputting the generated fragments into a Bayesian Classifier and applying cross-validation, they were able to predict quite accurately the potency of compounds from the test set (AUC values >0.85 were reported for 28 of the 36 kinases). Secondly, they were able to understand chemical space active against kinase targets better, showing that kinases unrelated by sequence were indeed able to bind similar ligand chemistry. However, while this effect was much less pronounced when looking at the overall similarity of ligands, it was much more significant when looking at the fragments on an individual basis. More recently, the group also described the design of novel kinase libraries with increased hit rates in prospective assays [42], and it is hoped that understanding kinase inhibitor selectivity better will lead to novel medicines with fewer side effects and better efficacy in vivo in the long run.

Extending this idea to "all" possible human protein targets, the question how proteins are related to another from the chemical side (in a similar way to the evolutionary relation from the biological, sequence side) is not only of academic interest, but it is indeed highly relevant also for the prediction of cross-reactivity of compounds, and hence the anticipation of adverse drug reactions. This type of analysis has only been made possible relatively recently with the advent of large bioactivity databases. Also Bayes methods were used in this field, exploiting the fact that the "model" which the Bayes Classifier generated is simply a

vector of conditional probabilities of features per activity class (This can also be interpreted as the relative frequency of a particular chemical feature, given that the compound is active against a particular target.). Those probability vectors can then be used to relate all bioactivity classes to each other, based on their similarity (calculated as a correlation coefficient between feature frequencies per bioactivity class) resulting from the Bayes' model. This approach has been applied in a recent study [43], using about 1,000 different bioactivity classes from the WOMBAT database as an input dataset. After generating Naïve Bayes bioactivity models, principal component analysis of the resulting space was performed. This analysis leads to approximately nine orthogonal axes in "bioactivity space" that were each defined by a set of distinct chemistry. Given the importance of GPCR ligands in pharmaceutical research (and hence also in databases such as WOMBAT), it is not surprising that the first of these axes was defined to a good extent by the chemistry of ligands belonging to this target class. While putting proteins into context via the ligand chemistry that binds to them encoded in a Bayes' model, it was also found that "Bayes Affinity Fingerprints," describing the similarity of a compound to each of the nine principal components, improved virtual screening retrieval results on a benchmark dataset derived from the MDDR database. Overall, understanding of bioactivity space across targets has the hope of contributing to the desired activity profile of ligands in the future; be they selective or multitarget drugs. More recently (and exchanging the Bayes model for expectation values stemming from the bioinformatics field), also prospective validations of the resulting protein target relationships have been published [44].

3.3. Postprocessing Docking Results

The prediction of ligand binding using fitting procedures into protein crystal structures or homology models, called docking, is an ubiquitous method in drug discovery. However, it has been repeatedly found that the prediction of binding energy (not so much the suggestion of a possible binding pose) is by no means trivial and an error-prone process [45, 46].

In order to improve the predictivity of docking results, it has been noticed recently [14] that the nature of docking results is actually not so different from the outcome of high-throughput screens, in the sense that in both sets of results output is highly enriched with noise, making the drawing of conclusions based on the data directly difficult. Hence, in the same way as HTS results are triaged, multiple groups investigated in which ways docking results can be postprocessed to give improved prediction results as to which compounds from a large database are active, and which ones are inactive. In one of the earliest studies of this type [21], the Bayes Classifier as well as PLS and inductive rules have been used to postprocess scoring results from seven different scoring

functions against four different targets (namely, the estrogen alpha receptor (ERalpha), matrix metalloprotease 3 (MMP3), factor Xa (fXa), and acetylcholine esterase (AChE)). All the resulting models, based on the seven-dimensional docking output scores, were multivariate models, hence also taking the mutual dependency of docking scores in a particular ligand–target combination into account. In this particular work, recursive partitioning was found to be superior to both PLS and Bayes methods; however, all of them were also outperforming one standard consensus scoring method, CScore. More recently also Binary QSAR has been evaluated on the above datasets using LUDI and MOE scoring functions [47], coming to roughly similar performance as the methods suggested by Jacobsson above.

More recently, also for the targets protein kinase B (PKB) and protein-tyrosine phosphatase 1B (PTP1B) Bayes postprocessing of docking results has been performed [48]. In this work, Dock, FlexX, and Glide scoring functions were used to dock compounds, and after finalizing the compound ranking, according to docking score, a rather stringent cutoff was applied to the scores to train a Bayesian Classifier on the best-ranking (assumed active) versus the worst-ranking (assumed inactive) compounds. The hypothesis was that the Bayesian Classifier would be able to both eliminate false-positives from the high-scoring areas of the list (since, on average, this particular kind of ligand chemistry was associated with rather lower scores) and also to recover false-negatives from the low-scoring areas of the list (since, in reverse, this particular kind of ligand chemistry was on average associated with rather high scores in the docking process). Note that this idea is actually related to identifying true-positive compounds from noisy screening data in pooled screens as discussed in the previous section [35] – in both cases the predicted (or measured) activity of a compound is put "into context" of the average activity this kind of chemistry exhibits, hence allowing for the detection and correction of unexpected events. And indeed, for all cases where initial enrichment of the ranked compound list from docking was observed (namely for all docking programs employed on PTP1B), actives could be enriched by postprocessing with the Naïve Bayesian Classifier quite significantly (such as in case of Dock from 4.5-fold enrichment in the top 10% of the database to 6.8-fold enrichment when using the Bayes Classifier). But also, vice versa, on PKB the ranked lists of both FlexX and Glide actually experience *derichment* of the number of active compounds upon applying the Bayes Classifier. Upon further analysis, this was found to be due to a large number of false positives at the extreme beginning of the hit list – unfortunately this was the region compounds were drawn from for training the activity model. Likewise, the combination of PKB and the Dock scoring function experiences derichment of actives already in the first docking step, rendering application of further

model training futile. Hence, while postprocessing of docking results can indeed greatly improve hit rates, care should be taken which kind of data is used to train the Bayes model, since it needs to drive compounds selection further into the right direction; otherwise, this step is doomed to fail. More recently, it has also been suggested that docking only part of a library is sufficient in order to train a Bayesian Model based on circular fingerprints on the docking actives and inactives for ranking the whole library, greatly reducing computational burden [49]. In this case the method was applied to CDK2 and estrogen receptor, and in this study in the order of 50 docking hits are needed in each case for generating a suitable activity model, here leading to 13-fold and 35-fold enrichment of the full compound set, respectively. In related work on CDK2 [50], rescoring schemes using both traditional consensus scoring as well as Bayesian Methods and linear discriminant analysis do come to a mixed result. While there is always a way to combine two scoring functions to outperform a single scoring function, both multivariate approaches and traditional consensus scoring seem to perform superior in particular cases.

4. The Bayes Classifier in Chemical Biology

4.1. Target Deconvolution and Merging Biochemical and Phenotypic Measurements

While the Bayes Classifier, as shown above, exhibits considerable performance in virtual screening tasks, where a novel ligand for a given target should be identified, also the reverse process can be performed, identifying a protein target for a given, orphan compound. This process, called "target prediction," has been extensively worked on at Novartis, originating from a paper of Nidhi et al. [51]. In this original work, compounds from the WOMBAT database were used to train the Naïve Bayesian Classifier in PipelinePilot, and subsequently compounds with solely phenotypic annotation data, such as "antihypertensives," were fed into the model. Target predictions for this set (as well as others) partly point to known targets such as the rennin-angiotensin system which confirms what is already known; but also ion channels and other proteins are predicted to be involved in decreasing blood pressure, raising the possibility of novel targets that can be tackled in the future. Later, integration with novel, information-rich screening techniques such as high-content screening (HCS) has been performed [52], where HCS gives purely a phenotypic output, but without hinting at a particular protein target or pathway responsible for this effect. Hence, integrating phenotypic screening [53] with ligand–target prediction tools can be used to characterize a compound more comprehensively, via its chemical structure, its effect on a biological system, as well as the phenotypic response it

causes in a biological system. While HCS is still in its infancy in many ways, regarding for example the standardization of assays and the analysis of the resulting data, given the turn in pharmaceutical industry to screen more and more in phenotypic assays, the deconvolution of targets will likely gain importance in the future.

While HCS readouts are one type of phenotypic readouts, adverse drug reactions are another example of this type, describing the response of a biological system to the application of a chemical. While some adverse reactions are understood, for the majority the preclinical profiling performed in industry [54] is still to a good extent based on personal experience and gut feeling. However, given that Bayes Classifiers can be trained on both ligand–target, as well as ligand–phenotypic effect relationships, bridging the space between phenotypic and mechanistic world is possible by comparing the conditional probability vectors originating from both model types to another. Thus, via the common language of chemistry, one can conclude from the similar chemistry present in on-target activities and datasets showing adverse reactions, which targets might be involved in adverse drug reactions (and which might, in turn, be suitable assays to be included in future preclinical profiling efforts). For data from WOMBAT and side effects annotated in the World Drug Index database this relation is shown in Fig. 2, with the most common side effects for the opioid μ receptor, the muscarinic m1 receptor, and cyclooxygenase-1 displayed (in white boxes). It can be seen that many side effects are well known in literature, such as dependence for opioid receptor agonists (such as heroine), dry mouth for m1 receptor agonists, and gastrointestinal bleeding for cyclooxygenase-1 inhibitors (such as aspirin). One can now of course go ahead and extend this model by taking many more adverse effects and many more on-target activities into account, elucidating new connections between both spaces and harnessing the true power of today's large chemogenomic databases.

4.2. Protein–Ligand Interactions

For many protein structures, despite the growth of structural biology archives such as the Protein Data Bank (PDB), until today no crystal or other structural models have been obtained. Still, in many cases a set of ligands binding to the target is known, hence allowing computational modeling to take place that suggests the most likely ligand–protein interactions around particular atoms of the ligand. In a truly comprehensive recent study, atoms were typed [55] according to a set of 30 predefined ligand atom types and 25 protein (target) atom types. Next, a selected set of 5,863 protein–ligand interaction files from the PDB were analyzed with respect to the frequency of encountering combinations of particular ligand and target atoms within a certain radius. In the generation of the Bayesian ligand–protein interaction model, the relative frequency of every ligand–protein interaction

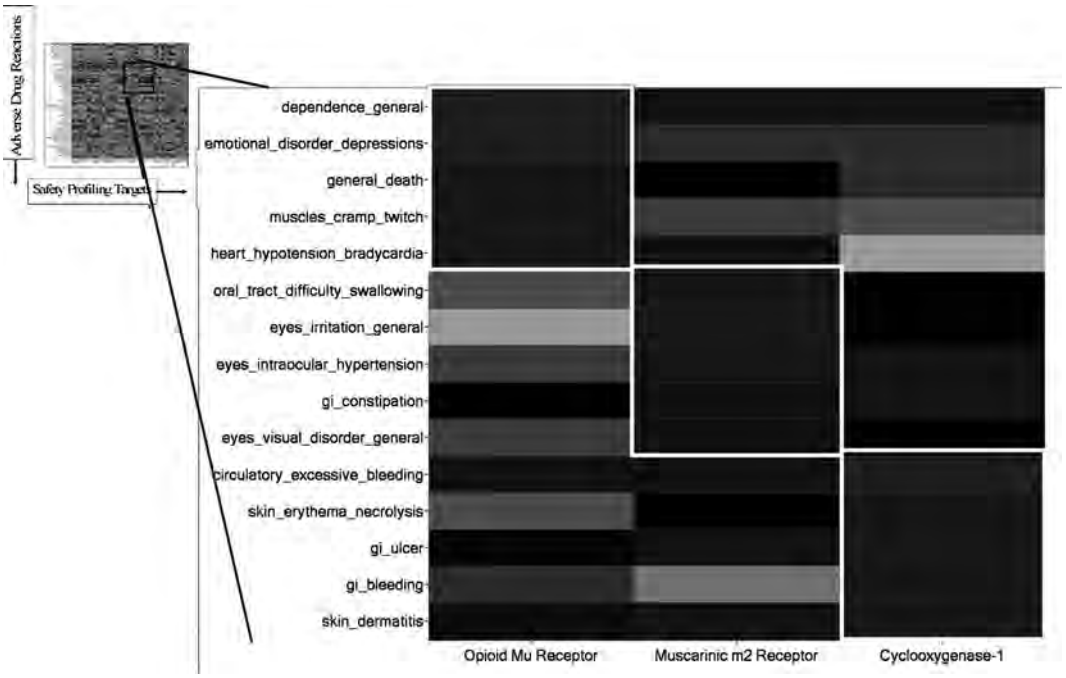


Fig. 2. Relationship between adverse drug reactions and on-target activities of drugs, analyzed by relating Bayes model from both areas. The most common side effects for the opioid μ receptor, the muscarinic m1 receptor, and cyclooxygenase-1 are shown here (surrounded by *white boxes*; for color figure *see* original reference in the text) and it can be seen that many side effects are well known in the literature, such as dependence for opioid receptor agonists (such as heroine), dry mouth for m1 receptor agonists, and gastrointestinal bleeding for cyclooxygenase-1 inhibitors (such as aspirin) are also seen in this dataset. The next step is to take more adverse effects and many more on-target activities into account, elucidating new connections between both space.

(with the respective atom types) was used as the prior in the Bayes formula, and spatial distributions of protein atoms around a ligand atom were modeled by Gaussian mixture models train using expectation maximization (EM) algorithms. After training, Bayes' formula can then be exploited; from the observed ligand–target interaction probabilities of particular atoms in the training set one can conclude, for a novel ligand of a protein, what the most likely interaction atoms in the protein will be. This can of course then be further exploited for the design of novel, potent ligands. Applications to ligands of Chlorella virus DNA ligase [56] and the ligand-binding domain of glutamate receptors [55] have been presented and, indeed, the predicted interactions in the protein conform to a good extent to the experimentally determined interactions. While the method is quite labor-intensive, it still has the potential for "structure-based ligand design in the absence of the structure," and hence prospective applications of it will be very interesting to see in the future.

5. Conclusions

Given their performance, versatility, and speed, Bayesian methods have truly deserved their place in the modeling toolbox in areas such as virtual screening and, due to their interpretability, also in areas such as chemical biology.

The Naïve Bayesian Classifier, despite its name and assumptions made in its derivation, shows often performance very close to iterative learning methods (such as support vector machines), while at the same time being a single-shot learner that requires significantly less time for model generation. In addition, in areas such as the analysis of HTS data, its tolerance to noise is beneficial and when postprocessing docking results, its application has repeatedly shown to increase hit rates in the resulting, prioritized set. Other Bayesian methods such as Bayesian Regularized Artificial Neural Networks make use of parameters to hold the model size at bay, alleviating overfitting problems previously encountered with standard feed-forward, back-propagation networks. Bayesian Inference Networks seem to improve performance in virtual screening; in addition they take the mutual dependency of compound features into account which is also appealing from the theoretical side, where very often distinct binding modes of different scaffolds are observed, each of which depends on a *distinct set* of ligand features necessary for binding.

In the area of chemical biology, the Bayes Classifier is able to bridge phenotypic and mechanistic bioactivity space, as shown for the areas of HCS and the analysis of targets involved in particular adverse drug reactions. Here, the easy interpretation of feature vectors, representing the relative frequencies of chemical features, makes bridging both spaces possible and gives insight into *why* particular functional groups are related to a phenotypic observation (adverse drug reaction, phenotypic screening readout, etc.). In addition, Bayes methods have been employed to give insight into ligand–protein interactions in a special manner, allowing potentially for "structure-based drug design in absence of a (protein) structure."

Overall, it can be hoped that data analysis methods based on Bayes' theorem will gain importance in the future, allowing pharmaceutical drug discovery to benefit from its main objective, namely "making the best decision in a given situation, given the data."

Acknowledgements

This work was supported by the Dutch Top Institute Pharma, project number: D1-105.

References

1. Bayes, T. (1763) An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London*, **53**, 370–418.
2. Kohavi, R., Becker, B., and Sommerfield, D. (1997) Improving simple Bayes. *Proc. 9th Europ. Conf. Mach. Learn.*, 78–87.
3. Domingos, P., and Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.*, **29**, 103–130.
4. Dougherty, J., Kovahi, R., and Sahami, M. (1995) Supervised and unsupervised discretization of continuous features. *Proc. 12th Int. Conf. Mach. Learn.*, 194–202.
5. Rish, I., Hellerstein, J., and Thathachar, J. (2001) An analysis of data characteristics that affect naive Bayes performance. *IBM Research Report RC21993.*
6. Rish, I., Hellerstein, J. L., and Jayram, T. S. (2001) An analysis of naive Bayes Classifer on low-entropy distributions. *IBM Research Report RC91994.*
7. Bender, A., and Glen, R. C. (2004) Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.*, **2**, 3204–3218.
8. Glick, M., Jenkins, J. L., Nettles, J. H., Hitchings, H., and Davies, J. W. (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive Bayesian classifiers. *J. Chem. Inf. Model.*, **46**, 193–200.
9. Lameijer, E. W., Kok, J. N., Back, T., and Ijzerman, A. P. (2006) Mining a chemical database for fragment co-occurrence: discovery of "chemical cliches". *J. Chem. Inf. Model.*, **46**, 553–562.
10. Abdo, A., and Salim, N. (2009) Similarity-based virtual screening with a Bayesian inference network. *ChemMedChem*, **4**, 210–218.
11. Cloutier, L. M., and Sirois, S. (2008) Bayesian versus Frequentist statistical modeling: a debate for hit selection from HTS campaigns. *Drug Discov. Today,* **13**, 536–542.
12. Zhou, Y. (2004) Choice of designs and doses for early phase trials. *Fundam. Clin. Pharmacol.*, **18**, 373–378.
13. Gilmore, S. J. (2008) Evaluating statistics in clinical trials: making the unintelligible intelligible. *Australas. J. Dermatol.*, **49**, 177–184; quiz 185–186.
14. Klon, A. E. (2009) Bayesian modeling in virtual high throughput screening. *Comb. Chem. High Throughput Screen.*, **12**, 469–483.
15. Labute, P. (1999) Binary QSAR: a new method for the determination of quantitative structure-activity relationships. *Pac. Symp. Biocomput.*, **4**, 444–455.
16. Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. Q., et al. (2007) Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided Mol. Des.*, **21**, 53–62.
17. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.*, **44**, 1708–1718.
18. Gao, H., Williams, C., Labute, P., and Bajorath, J. (1999) Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.*, **39**, 164–168.
19. Stahura, F. L., Godden, J. W., Xue, L., and Bajorath, J. (2000) Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.*, **40**, 1245–1252.
20. Labute, P., Nilar, S., and Williams, C. (2002) A probabilistic approach to high throughput drug discovery. *Comb. Chem. High Throughput Screen.*, **5**, 135–145.
21. Jacobsson, M., Liden, P., Stjernschantz, E., Bostrom, H., and Norinder, U. (2003) Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.*, **46**, 5781–5789.
22. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.*, **44**, 170–178.
23. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.*, **44**, 1177–1185.
24. Glen, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., and Smith, J. (2006) Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs*, **9**, 199–204.
25. Bender, A., Mussa, H. Y., Gill, G. S., and Glen, R. C. (2004) Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.*, **47**, 6569–6583.

26. Liu, Y. (2004) A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.*, **44**, 1823–1828.
27. Godden, J. W. and Bajorath, J. (2003) An information-theoretic approach to descriptor selection for database profiling and QSAR modeling. *QSAR Comb. Sci.*, **22**, 487–497.
28. Vogt, M., and Bajorath, J. (2008) Bayesian similarity searching in high-dimensional descriptor spaces combined with Kullback-Leibler descriptor divergence analysis. *J. Chem. Inf. Model.*, **48**, 247–255.
29. Diller, D. J., and Hobbs, D. W. (2004) Deriving knowledge through data mining high-throughput screening data. *J. Med. Chem.*, **47**, 6373–6383.
30. Wasserman, L. (2000) Bayesian model selection and model averaging. *J. Math. Psychol.*, **44**, 92–107.
31. Angelopoulos, N., Hadjiprocopis, A., and Walkinshaw, M. D. (2009) Bayesian model averaging for ligand discovery. *J. Chem. Inf. Model.*, **49**, 1547–1557.
32. Parker, C. N. (2005) McMaster university data-mining and docking competition – computational models on the catwalk. *J. Biomol. Screen.*, **10**, 647–648.
33. Rogers, D., Brown, R. D., and Hahn, M. (2005) Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.*, **10**, 682–686.
34. Bender, A., Mussa, H. Y., and Glen, R. C. (2005) Screening for dihydrofolate reductase inhibitors using MOLPRINT 2D, a fast fragment-based method employing the naive Bayesian classifier: limitations of the descriptor and the importance of balanced chemistry in training and test sets. *J. Biomol. Screen.*, **10**, 658–666.
35. Glick, M., Klon, A. E., Acklin, P., and Davies, J. W. (2004) Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screen.*, **9**, 32–36.
36. Schneider, G., Schneider, P., and Renner, S. (2006) Scaffold-hopping: how far can you jump? *QSAR Comb. Sci.*, **25**, 1162–1171.
37. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2004) Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.*, **2**, 3256–3266.
38. Crisman, T. J., Bender, A., Milik, M., Jenkins, J. L., Scheiber, J., Sukuru, S. C., et al. (2008) "Virtual fragment linking": an approach to identify potent binders from low affinity fragment hits. *J. Med. Chem.*, **51**, 2481–2491.
39. Burden, F. R., and Winkler, D. A. (1999) Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.*, **42**, 3183–3187.
40. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
41. Sutherland, J. J., Higgs, R. E., Watson, I., and Vieth, M. (2008) Chemical fragments as foundations for understanding target space and activity prediction. *J. Med. Chem.*, **51**, 2689–2700.
42. Vieth, M., Erickson, J., Wang, J., Webster, Y., Mader, M., Higgs, R., et al. (2009) Kinase inhibitor data modeling and de novo inhibitor design with fragment approaches. *J. Med. Chem.*, **52**, 6456–6466.
43. Bender, A., Jenkins, J. L., Glick, M., Deng, Z., Nettles, J. H., and Davies, J. W. (2006) "Bayes Affinity Fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multi-target drugs a feasible concept? *J. Chem. Inf. Model.*, **46**, 2445–2456.
44. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.
45. Glen, R. C., and Allen, S. C. (2003) Ligand-protein docking: cancer research at the interface between biology and chemistry. *Curr. Med. Chem.*, **10**, 767–782.
46. Warren, G. L., Andrews, C. W., Capelli, A.-M., Clarke, B., LaLonde, J., Lambert, M. H., et al. (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, **49**, 5912–5931.
47. Prathipati, P., and Saxena, A. K. (2006) Evaluation of binary QSAR models derived from LUDI and MOE scoring functions for structure based virtual screening. *J. Chem Inf. Model.*, **46**, 39–51.
48. Klon, A. E., Glick, M., Thoma, M., Acklin, P., and Davies, J. W. (2004) Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. *J. Med. Chem.*, **47**, 2743–2749.
49. Yoon, S., Smellie, A., Hartsough, D., and Filikov, A. (2005) Surrogate docking: structure-based virtual screening at high throughput speed. *J. Comput. Aided Mol. Des.*, **19**, 483–497.
50. Cotesta, S., Giordanetto, F., Trosset, J. Y., Crivori, P., Kroemer, R. T., Stouten, P. F., et al. (2005) Virtual screening to enrich a compound collection with CDK2 inhibitors

using docking, scoring, and composite scoring models. *Proteins*, **60**, 629–643.
51. Nidhi, Glick, M., Davies, J. W., and Jenkins, J. L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.*, **46**, 1124–1133.
52. Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G. W., Tao, C. Y., et al. (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat. Chem. Biol.*, **4**, 59–68.
53. Feng, Y., Mitchison, T. J., Bender, A., Young, D. W., and Tallarico, J. A. (2009) Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat. Rev. Drug Discov.*, **8**, 567–578.
54. Whitebread, S., Hamon, J., Bojanic, D., and Urban, L. (2005) In vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov. Today*, **10**, 1421–1433.
55. Rantanen, V. V., Gyllenberg, M., Koski, T., and Johnson, M. S. (2003) A Bayesian molecular interaction library. *J. Comput. Aided Mol. Des.*, **17**, 435–461.
56. Rantanen, V. V., Denessiouk, K. A., Gyllenberg, M., Koski, T., and Johnson, M. S. (2001) A fragment library based on Gaussian mixtures predicting favorable molecular interactions. *J. Mol. Biol.*, **313**, 197–214.

Chapter 8

Reduced Graphs and Their Applications in Chemoinformatics

Kristian Birchall and Valerie J. Gillet

Abstract

Reduced graphs provide summary representations of chemical structures by collapsing groups of connected atoms into single nodes while preserving the topology of the original structures. This chapter reviews the extensive work that has been carried out on reduced graphs at The University of Sheffield and includes discussion of their application to the representation and search of Markush structures in patents, the varied approaches that have been implemented for similarity searching, their use in cluster representation, the different ways in which they have been applied to extract structure–activity relationships and their use in encoding bioisosteres.

Key words: Reduced graph, Graph reduction, Similarity searching, Bioisosterism, Structure–activity relationships, Markush structures, Generic structures, Database search

1. Introduction

Reduced graphs [1] provide summary or abstract representations of chemical structures and are generated by collapsing connected atoms into single nodes, edges are then formed between the nodes according to bonds in the original structure. Reduced graphs have been used in a variety of applications in chemoinformatics ranging from the representation and search of Markush structures in chemical patents to the identification of structure–activity relationships (SARs). Many different graph reduction schemes have been devised, and the optimal scheme is likely to depend on the particular application. Examples of different types of reduced graphs are shown in Fig. 1. The idea of characterising chemical structures by their structural components is long established in chemical information and is implicit in most systematic chemical nomenclature: structures are fragmented into

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_8, © Springer Science+Business Media, LLC 2011

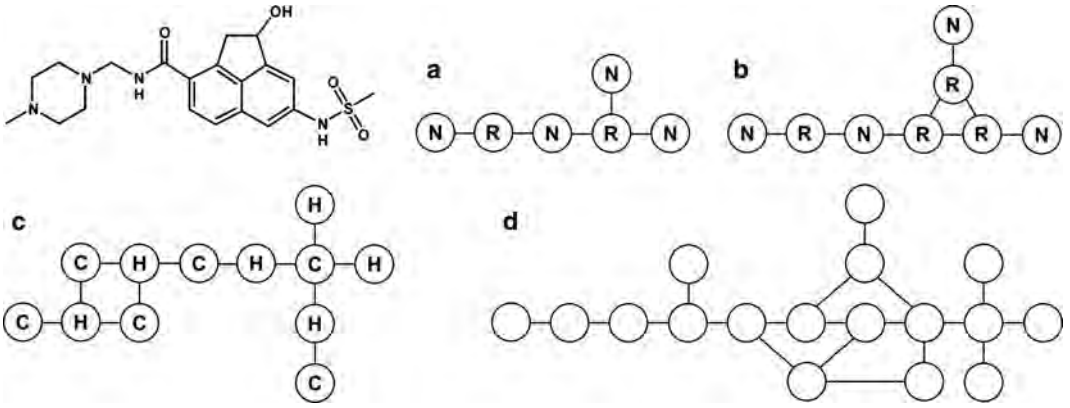


Fig. 1. Different graph reduction schemes. (**a**) A ring/non-ring reduction where a fused ring system is reduced to a single node. (**b**) A ring/non-ring reduction where each smallest ring is treated as an individual node. (**c**) A carbon/heteroatom reduction. (**d**) A homeomorphic reduction in which atoms of degree two are removed. The node types are denoted as follows: *R* ring, *N* non-ring, *C* carbon, and *H* heteroatom.

ring systems and acyclic components, which are described individually with conventions used to indicate how they are connected, for example, *N*-(4-hydroxyphenyl) acetamide (the systematic name for paracetamol). Reduced graphs also aim to summarise structures according to their structural components, however in contrast to nomenclature systems, they retain structural information on how the components are connected in graphical form. This encoding of topology enables structural comparisons to be made, which cannot be achieved through the use of nomenclature.

In this chapter, we focus on the extensive work that has been carried out on reduced graphs at The University of Sheffield for a variety of different applications. We also recognise the substantial efforts made by other groups in related methods, notably the feature trees approach by Rarey et al. [2, 3] and the extended reduced graph, ErG, by Stiefl et al. [4, 5], and provide a brief summary of these approaches.

2. Reduced Graphs for Searching Markush Structures

Reduced graphs were first used at Sheffield as a component of a search system for Markush structures [1, 6]. Markush structures (also known as generic structures) are chemical structures that involve the specification of lists of alternative substituents attached to a central core structure. They occur frequently in chemical patents where they are used to describe a large and often unlimited number of structures with the aim of protecting a whole class of compounds rather than a few specific examples. An example Markush structure is shown in Fig. 2 and consists of a central core

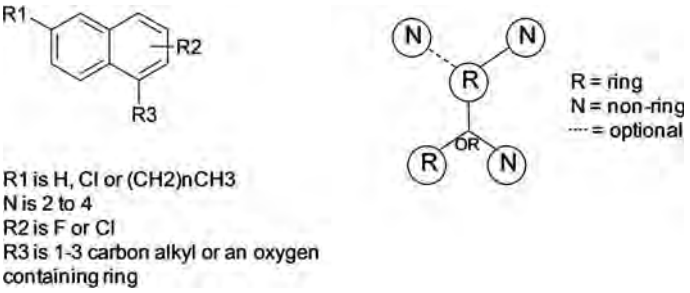


Fig. 2. A Markush structure and its reduced graph representation based on a ring/non-ring reduction scheme.

group with variable R-groups that are used to represent lists of alternative substituents (or substructures) attached to the core. Markush structures pose several difficulties for storage and retrieval. In addition to handling the large number of compounds encoded in a single representation and dealing with different ways of partitioning a structure into substructures, one of the major difficulties is the use of generic nomenclature to indicate that a substituent may be any member of a homologous series, for example, in expressions such as "R1 is an alkyl group". Generic nomenclature presents difficulties for search since it is necessary to be able to match specific instances of a homologous series with the generic term, for example, to recognise that "methyl" is an instance of "alkyl".

Reduced graphs were developed in the Sheffield Generic Chemical Structures Project [7] to provide an additional level of search that is intermediate in complexity between the traditional fragment screening and atom-by-atom search methods that were developed for specific structures, and to provide an effective way of dealing with generic nomenclature [6]. In the Sheffield project, homologous series are represented by parameter lists which indicate the structural features that characterise the series such as: the number and type of rings present and the presence or absence of heteroatoms, etc. Most of the substituents that are expressed as homologous series in patents can be classified as ring or non-ring (for example, aryl, heterocycle, alkyl, alkene, etc.) and can therefore be represented as single nodes in a ring/non-ring (R/N) graph reduction scheme. The reduced graph representation of the generic structure is also shown in Fig. 2; the reduced graph is rooted on the central ring node, which is derived from the core structure and contains alternative nodes indicated by the branched edge labelled "OR" and an optional node indicated by the dashed edge. In the example shown, the partitioning of the generic structure into partial structures corresponds with the node definitions. However, in other cases, a single reduced graph node might span different partial structures in the generic structure.

The searching of Markush structures is carried out at three levels. The first level is a fragmentation search in which fragments are generated from both the structural fragments and the parameter lists used to represent the generic nomenclature: the fragments are organised as those which MUST be present in the generic structure and those that MAY be present since they occur in alternative substructures. The reduced graph search is based on graph matching procedures and is considerably faster than graph matching at the atom and bond level due to the relatively small size of the reduced graphs. The final search is an atom-by-atom search modified to deal with the generic nomenclature. The three search methods are applied in sequence: for a given query, those database compounds that pass the fragment stage are passed to the reduced graph search and finally those compounds remaining after the reduced graph search are subjected to the most time-consuming atom-by-atom search. Although the Sheffield search system did not become a public system in its own right, it undoubtedly had a major influence on the Markush DARC system of Derwent Information Limited [8] and the MARPAT system of Chemical Abstracts Service [9, 10].

3. Reduced Graphs for Similarity Searching

Since the advent of similarity searching in the 1980s, much effort has been expended on developing new descriptors with the aim of identifying compounds that share the same activity. The first similarity searching procedures were developed using fragment bitstrings that were devised for substructure search [11, 12]. These proved to be remarkably successful although this good performance was, in part, due to the nature of the datasets on which they were evaluated, which often consisted of series of structural analogues. A more recent focus in similarity searching has been the identification of compounds exhibiting the same activity but belonging to different lead series, a technique that has become known as scaffold hopping [13]. Such compounds offer important advantages over structural analogues: there is the potential to move away from the patent space of the query compound; and they provide the possibility of exploring more than one lead in parallel, with clear advantages should one series fail due to poor ADME properties or difficult chemistry.

Various graph reduction schemes have been developed for similarity searching. In this context the challenge is to reduce structures so that their pharmacophoric features are highlighted to enable compounds that share the same activity but belong to different chemical series to be perceived as similar. Figure 3 shows a series of compounds that are active at opioid receptors. The

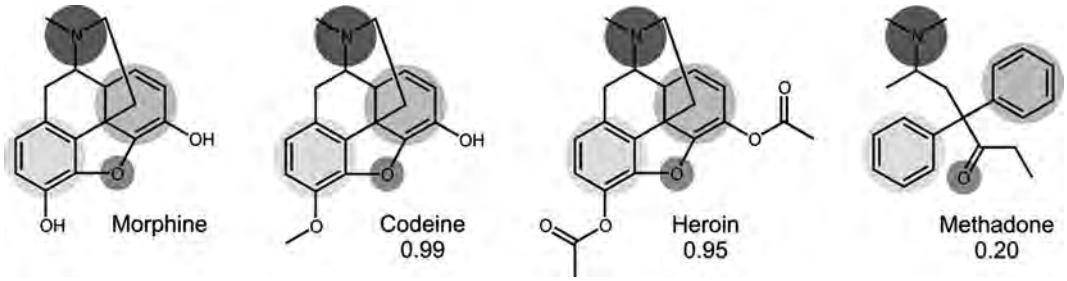


Fig. 3. The similarities of codeine, heroin and methadone are shown to morphine based on Daylight fingerprints and the Tanimoto coefficient.

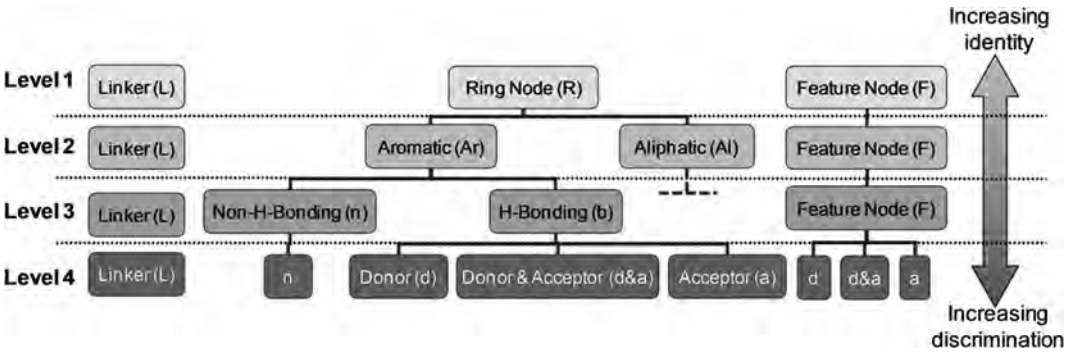


Fig. 4. A hierarchy of reduced graphs.

similarities of each of codeine, heroin and methadone to morphine are shown based on Daylight fingerprints [14] (a conventional 2D fingerprint) and the Tanimoto coefficient. The obvious 2D structural similarities of codeine and heroin to morphine are reflected in the high scores. However, methadone scores poorly despite having similar activity. The shaded spheres indicate a mapping between the structures that is based on their common functional groups and reveals similarities between methadone and the other compounds, which are not evident using conventional 2D fingerprints. When used for similarity searching, the aim of the graph reduction approach is to recognise such mappings so that the resulting reduced graphs can be thought of as topological pharmacophores.

4. Varying the Level of Specificity

Different graph reduction and node labelling schemes have been devised that vary in the level of specificity that is encoded and therefore in the degree of discrimination that is achieved between different structures. Figure 4 shows four levels of node specificity

for a reduction scheme based on three node types: Rings, Features and Linkers. In this scheme, linkers and features are distinguished using the concept of isolated carbons, which are acyclic carbon atoms that are not doubly or triply bonded to a heteroatom [14]. Connected isolated carbon atoms form linker nodes with the remaining connected acyclic components defining feature nodes. Non-hydrogen bonding terminal atoms are removed as indicated by the exclusion of the terminal methyl groups in structure A, Fig. 5. The different levels in the hierarchy are derived by further describing the nodes according to the properties of their constituent atoms in terms of aromaticity and hydrogen bonding character. As the level of detail encoded within the nodes is increased, the number of unique reduced graphs that are represented in a database increases, see Fig. 5. In experiments on the World Drug Index database, Gillet et al. determined that reduced graphs at level four in the hierarchy were most effective in discriminating between actives and inactives [15].

Variations on this basic approach have since been described, which include the definition of additional node types such as positively and negatively ionisable groups. Flexibility in the definition of node types is generally achieved through the use of user-defined SMARTS definitions for various groups such as hydrogen bond donors and acceptors.

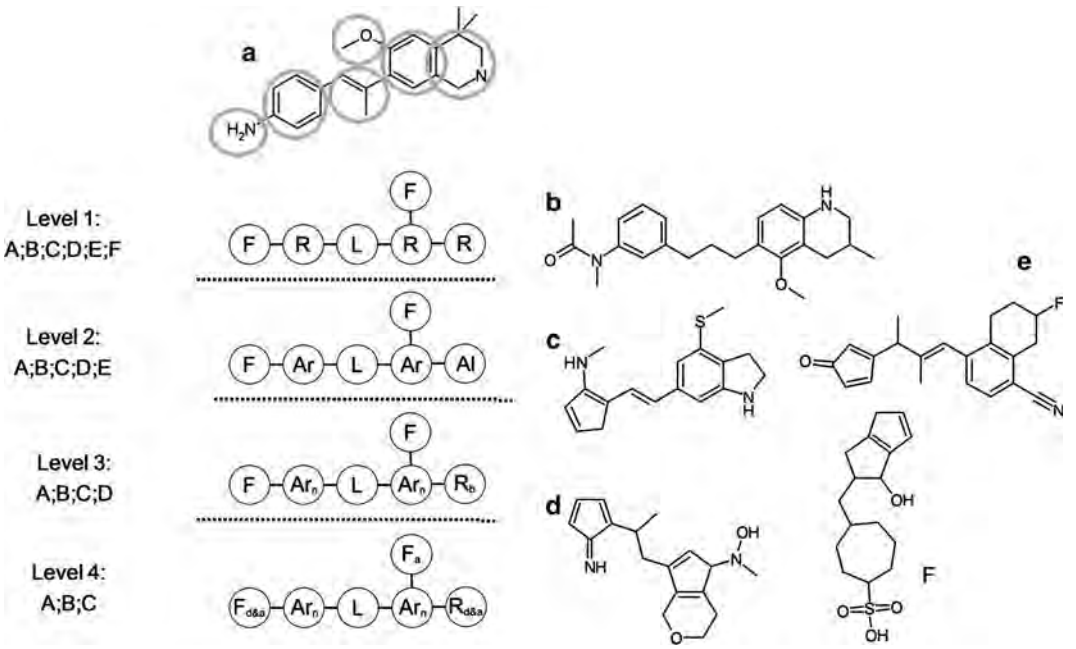


Fig. 5. The reduced graph for compound (**a**) at each level in the hierarchy in Fig. 4 is shown together with a series of related compounds: (**b–f**). At level one, all compounds are represented by the same reduced graph, at level 2, compounds A–E share the same reduced graph through to level 4 where only compounds (**a–c**) share the same reduced graph. The discrimination between structures is dependent on the level of descriptions encoded within the reduced graph.

5. Comparing Reduced Graphs Using Fingerprints

Various approaches have been devised to enable the similarity between a pair of molecules to be calculated based on their reduced graph representations. In analogy with the use of fragment bitstrings to compare chemical graphs, a similar approach has been taken to represent reduced graphs as binary vectors. For example, a mapping of node types to atoms not in the usual organic set, such as the transition metals, allows the reduced graphs to be represented as SMILES strings, as shown in Fig. 6, and the Daylight fingerprinting routines to be used to generate path-based fingerprints from reduced graphs [15]. While this approach provided a convenient way of comparing reduced graphs, the different characteristics of reduced graphs, relative to the structures from which they are derived, are such that the resulting fingerprints are suboptimal for quantifying the similarity between reduced graphs. For example, reduced graphs consist of fewer nodes than their corresponding chemical graphs so that the resulting fingerprints can be quite sparse and small changes in a chemical structure, such as the insertion of a heteroatom into an acyclic chain, can result in a quite different set of nodes and therefore fingerprint.

Improved performance was obtained by representing the reduced graphs as node-pair descriptors [16], which are similar in concept to the more familiar atom-pair descriptors developed by Carhart et al. [11]. For example, Harper and colleagues developed fingerprints based on node-edge pairs in which additional bits are set, for example, to encode branch points so that more of the topology of the reduced graph is represented and to encode paths of length one shorter than the actual length to introduce a fuzziness to the fingerprint [17]. The "fuzzy bits" enable the similarity between RGs that differ by the insertion or deletion of a single node to be perceived, which would otherwise give rise to a set of node-edge pairs of different lengths.

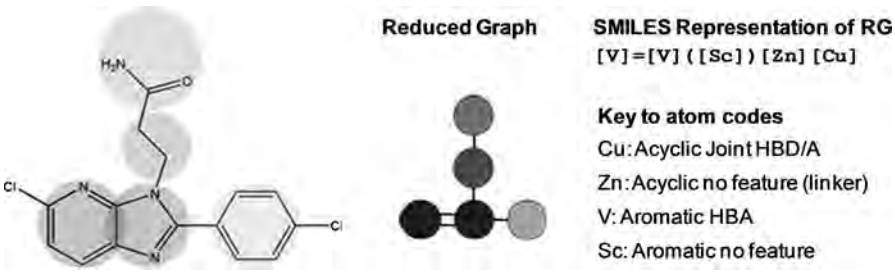


Fig. 6. A reduced graph represented as a SMILES string. Note that terminal, non-hydrogen bonding atoms have been removed when forming the reduced graph and that the fused ring nodes are represented by the "=" symbol.

Harper et al. also developed an edit distance method to quantify the similarity between reduced graphs, which is based on the cost of converting one reduced graph to the other by considering mutation, insertion and deletion of nodes. The edit distance technique is well known in computational biology where it is used for sequence comparisons with similarity related to the number of operations required to change one sequence to another. In the context of reduced graphs, edit distance is well suited to dealing with the problem of small changes in chemical structure leading to different patterns of nodes, for example, by the insertion of a heteroatom into a carbon chain. Furthermore, different weights can be assigned to different node operations to reflect similarities in node types. For example, in Harper's work the substitution of a "donor" to a "donor and acceptor" node was assigned a low cost, whereas the substitution of a "donor" to a "negatively ionisable group" was assigned a high cost. Harper showed that combining the edit distance similarity measure with a node-pair fingerprinting method improved the performance of the reduced graphs in similarity searches compared to the path-based fingerprints. The edit distance method is illustrated in Fig. 7: the left hand side shows the minimum cost of converting reduced graph B into A based on the matrix of substitution costs and the insertion/deletion costs shown on the right.

The costs assigned to the individual node operations by Harper were based on intuition. Subsequently, Birchall et al. [18] used a genetic algorithm to identify optimised sets of weights that gave improved performance over a variety of activity classes extracted from the MDL/Symyx Drug Data Report (MDDR) database [19]. They also generated sets of weights optimised on specific activity classes and showed that class-specific weights could

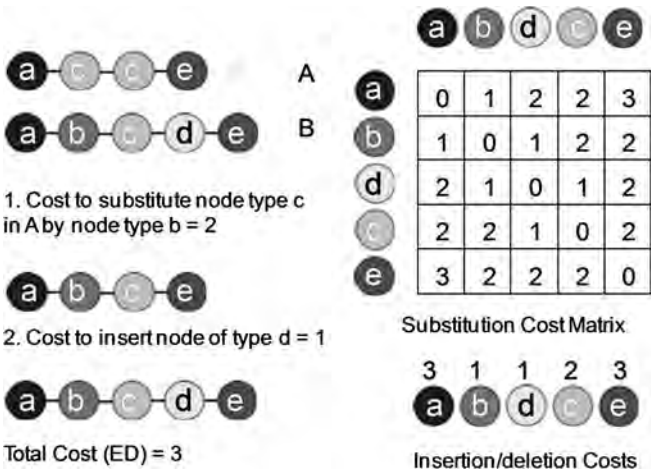


Fig. 7. The edit distance cost of converting the pattern of nodes in A to B is 3 based on the substitution cost matrix and insertion/deletion costs shown on the *right*.

not only improve retrieval performance but could also provide some clues on the underlying structure–activity relationship.

6. Comparing Reduced Graphs Using Graph Matching Procedures

By definition, reduced graphs contain fewer nodes and edges than the chemical graphs from which they are derived, making them more amenable to graph matching procedures. Takahashi et al. described an early approach to the use of graph matching techniques to compare reduced graph representations, albeit based on a very small number of compounds [20]. They considered a set of five structurally diverse antihistamines and a set of six antipsychotropic agents, and in both cases, some of the structural similarities were found. In more recent work, Barker et al. represented the reduced graph as a fully connected graph in which the edges represent bond distances in the original chemical graph and used maximum common subgraph (MCS) techniques to calculate the similarity between pairs of reduced graphs using much larger datasets [21]. They demonstrated improved performance of the reduced graph relative to Daylight fingerprints both in terms of the recall of actives and in the diversity of the actives retrieved thus suggesting that reduced graphs might be beneficial in scaffold-hopping applications.

7. Clustering

The reduced graph approach has also been used for various clustering applications. Clustering is widely used to present sets of compounds to chemists for review, for example, typically the results from a high-throughput screening exercise will be clustered and clusters that are enriched in active compounds will be examined in an attempt to extract structure–activity information. The most commonly used clustering techniques are based on traditional 2D fingerprints that are derived from the chemical structures themselves; however, when using such fingerprints, it may be difficult to decipher the structural commonalities that are present within a cluster. Harper et al. used reduced graphs to cluster high-throughput screening data [17]. Each molecule is represented by several *motifs,* which include the reduced graph, near neighbours of the reduced graph in which single nodes are deleted or changed, and Bemis and Murcko frameworks [22]. Molecules that share a common motif are clustered together and the clusters are sorted with large clusters consisting predominantly of active compounds being presented to the user first. The

reduced graphs and frameworks allow the structural characteristics of the compounds to be easily seen, in contrast to clustering based on conventional fingerprints.

In related work, Gardiner et al. have used reduced graphs to identify cluster representatives [23]. Here a dataset is clustered using conventional 2D fingerprints, the members of a cluster are then represented as reduced graphs and an MCS algorithm is applied iteratively in order to obtain one or more reduced graph cluster representatives. The reduced graphs offer two advantages for this application: first, their small size means that the MCS comparisons can be run in real-time; and second, the cluster representatives can be mapped back to the original structures that they represent, allowing the chemists to interpret the key functionalities required for activity. The method also enables multiple series present within the same cluster to be identified as well as related clusters by comparing representatives from different clusters.

8. Reduced Graphs for Identifying SARs

Reduced graphs have been used in conjunction with recursive partitioning in order to derive structure–activity relationship models. As proof of principle, an SAR model was developed for angiotensin II receptor antagonists and compared with the known literature [16]. A fingerprint representation of the reduced graph was used to determine the splitting criteria in a decision tree based on a training set of 100 actives and 2,000 inactives extracted from the MDDR database [19]. A portion of the resulting tree is shown in Fig. 8 with the shaded box highly enriched in actives and

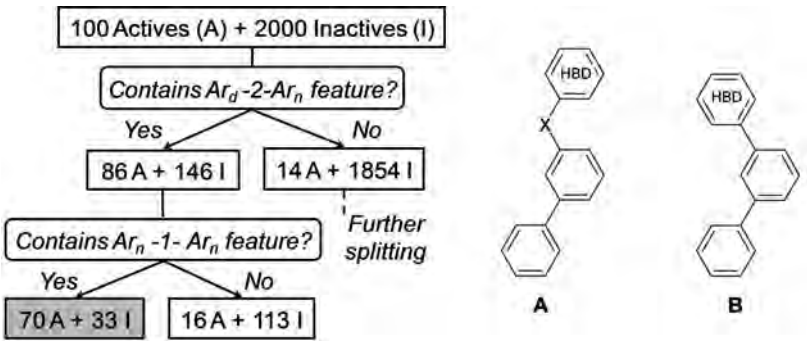


Fig. 8. A decision tree generated for angiotensin II antagonists based on reduced-graph representations. Substructure A is consistent with the node-edge pairs in the *shaded box* and is consistent with the known SAR. However, a limitation of the use of node-edge pairs for this application is that these two node-edges pairs are also present in other substructures, such as, B which may not be relevant to the SAR.

containing 70 of the active compounds. The splits in the tree are based on the presence/absence of node-edge pairs: Ar_d-2-Ar_n represents an aromatic ring containing a hydrogen bond donor separated by two edges from an aromatic ring with no hydrogen bonding character; Ar_n-1-Ar_n represents two aromatic rings with no hydrogen bonding characteristics separated by a single edge. These two node-edge pairs can be combined to represent the substructure A, which compares well with the 2D SAR model for angiotensin II receptor antagonists described by Bradbury et al. [24]. The approach was subsequently used in a procedure to select compounds for screening against a kinase inhibition assay with a hit rate of around 7% reported.

A disadvantage of the use of fingerprint representations to represent SARs is the loss of information on how the node-edge pairs are connected. For example, substructure A in Fig. 8 represents one way in which the node-edge pairs could be combined; however, there are other arrangements of rings that are also consistent with the same set of node-edge pairs, for example, substructure B. More recently, Birchall et al. have developed an evolutionary algorithm (EA) to grow reduced graph queries (subgraphs) with the aim of discriminating between actives and inactives in high-throughput screening data [25]. The reduced graph queries are encoded as SMARTs strings (such as that shown in Fig. 9) and allow a more detailed description of the structure–activity relationship to be developed. For example, a query can consist of any number of connected (or even disconnected nodes). Moreover, the use of atom primitives in the SMARTs language (such as OR and NOT logic) enables the range of substructures

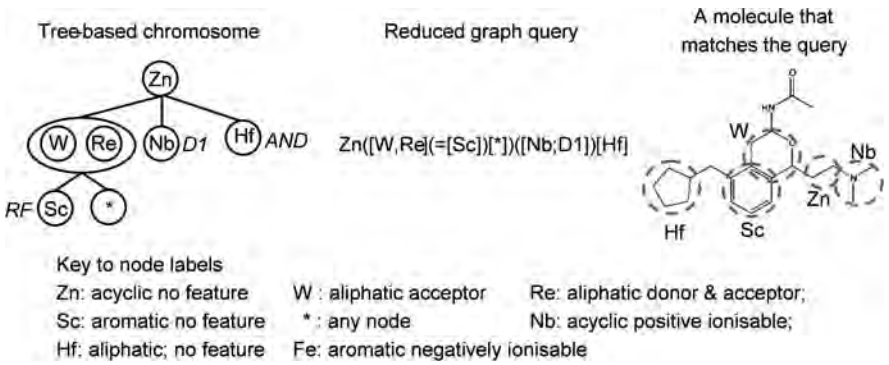


Fig. 9. A reduced graph query is shown as a SMARTS string in the *centre*. The *left-hand side* shows how the SMARTS string is mapped to a tree-based chromosome. The SMARTS primitives are tagged to nodes in the chromosome: D1 indicates degree 1; AND indicates a disconnected node (shown as "." in the SMARTS); RF indicates a ring fusion which is represented by a double bond in the SMARTS string. Two nodes are grouped to indicate that they represent alternative nodes. The *right-hand side* shows a molecule that matches the query with the nodes corresponding to the query highlighted.

that can be captured in a single expression to be extended. For example, a series of alternative node types can be specified at a given location in a subgraph to allow expressions such as "non-feature ring node OR acceptor ring node". The SMARTS expressions are mapped to tree-based chromosomes with the primitives tagged to nodes as shown in Fig. 9. Tree-based evolutionary operators have been developed to enable new trees to be evolved through the exchange of subtrees between chromosomes and various mutation operators.

A chromosome is evaluated by parsing the tree to generate a SMARTS query which is then searched across a training set of actives and inactives, also represented as reduced graphs. Fitness is measured using the *F*-measure, which is the harmonic mean of precision (P), the ratio of actives to total compounds retrieved and recall (R), the fraction of the actives retrieved, as follows: $F = 2PR/(P + R)$. The EA has been configured to evolve reduced graph queries that maximise the *F*-measure.

When applied to various activity classes extracted from the MDDR database [19], the EA was able to evolve reduced graph queries that give good classification rates and which encode structure–activity information that is readily interpreted by chemists. The approach was subsequently extended, first, to explore trade-offs in recall and precision and, second, to allow multiple SARs to be extracted from a single activity class [26]. The rationale for exploring the trade-off between precision and recall is that the optimum balance between these two objectives may depend on the application. For example, when seeking a structure–activity model it may be of interest to evolve a query with high precision at the expense of relatively low recall. Conversely, when evolving a query to be used in virtual screening, it may be more appropriate to choose a query that has higher recall but lower precision or even to choose a query that returns the same number of hits as the screening capacity. By treating recall and precision as independent objectives in a multiobjective optimisation procedure, a range of solutions are found which vary from high recall-low precision queries through to low recall-high precision queries. Multiple queries are evolved through the introduction of a third objective, called uniqueness, which compares each query with all others in the population. A query receives a high uniqueness score if the actives that it retrieves are not found by other queries in the population. This extension enables multiple SARs to be derived where each SAR describes a different set of active compounds. The combination of these complementary SARs allows for improved recall and precision as well as increasing the level of detail in the overall SAR description of a given activity class.

9. Reduced Graphs for Encoding Bioisosteres

Bioisosteres are structural fragments that can be exchanged without significant change to a molecule's biological activity. Since bioisosteres may be quite different in structure, e.g. tetrazole and carboxylic acid, it is challenging for conventional similarity measures to reflect their functional similarity. Graph reduction approaches are an attractive means of dealing with such equivalences as they allow several different structures to be encoded as the same node type. Birchall et al. [27] investigated how bioisostere information could be exploited in similarity searching using a graph-matching approach. Bioisosteres extracted from the BIOSTER database [28] were often found to be encoded by the same node type, supporting the applicability of the reduced graph encoding. However, there were also many cases where the bioisosteric fragments were not encoded as the same node type or even by the same number of nodes. The graph reduction and matching schemes were then modified to recognise and permit matches between instances of bioisosteric fragments, enhancing the similarity between molecules containing such fragments. Similarity searches in the WOMBAT database [29] found that although this approach clearly demonstrates scaffold hopping potential there is a significant trade-off in terms of the number of inactives that are also retrieved. The issue here arises from the fact that bioisosteric equivalences are often dependent on the specific context in which they are considered, both in terms of the intra-molecular environment and the extra-molecular environment, something that is perhaps too complex for broad generalisation based on the available data. By altering the rules used for graph partitioning, node type assignment and node type matching, reduced graphs provide the flexibility to allow the recognition of increasingly structurally distinct equivalences. However, this must be balanced against the degree of information loss inherent in graph reduction that may lead to the recognition of unreasonable equivalences. The key is in deciding what constitutes a reasonable equivalence.

10. Related Approaches

The intention of this chapter has been to summarise the extensive work carried out on reduced graphs at Sheffield, however, in acknowledgement of the significant contributions made by other groups, we briefly summarise the closely related approaches of feature trees and ErG. The feature tree, developed by Rarey and Dixon, also seeks to generalise chemical structures by emphasising

their functional features [2]. A ring/non-ring reduction similar to that in Fig. 1a is carried out except that a separate node is assigned to each non-terminal acyclic atom. The resulting structure is a tree (i.e. it does not contain any cycles), which allows significant improvements in speed when comparing two feature trees due to the greater efficiency of tree-matching algorithms relative to graph-matching. Each node in the tree is "labelled" with a range of *features* derived from its constituent atom(s) such as their volume and molecular interaction capabilities. Calculating the similarity between two trees is based on first finding a match of sub-trees and then using a weighted combination of the feature similarities of the matching nodes. Feature trees of a lower specificity can be derived by collapsing sub-trees into single nodes to give rise to a hierarchy of representations, which allows similarities to be determined at varying levels of specificity. Feature trees have been applied to a number of applications including: similarity searching based on a single query [2], similarity searching based on multiple queries by combining the queries into a multiple feature tree model called MTree [30] and fast similarity searching in very large combinatorial libraries [3].

The extended reduced graph (ErG) approach developed by Stiefl et al. [4] is similar to the reduced graph but includes a number of extensions. For example, charged and hydrophobic features are encoded explicitly and rings are encoded as ring centroids with substituted ring atoms encoded as separate nodes. The nodes are connected according to the shortest paths in the chemical graph. Although the ErG is a more complex graph than the reduced graph, positional information is better conserved and inter-feature distances in the original molecule tend to be more accurately represented. Furthermore, separation of the ring features from the ring itself permits similarity to be reflected between rings of different feature types. For example, while the reduced graph encodes pyrrole and phenyl rings as different node types, the ErG approach represents pyrrole as an aromatic node joined to a donor node, which retains some commonality with the single aromatic node resulting from a phenyl group. The ErG can be encoded in a fingerprint, similar to those developed for reduced graphs; however, Stiefl et al. use a hologram approach where each bit encodes a count of the fragment frequency rather than binary presence or absence. Some fuzziness in matching is permitted by incrementing the bits for paths that are both longer and shorter than the path length. In simulated virtual screening experiments across a range of activity classes, ErG was found to be comparable to Daylight fingerprints of chemical graphs, in terms of enrichments; however, they were found to be more effective for scaffold hopping since a greater diversity of structural classes were found. Stiefl and Zaliani [5] also describe an extension of ErG in which a weighting scheme is used to increase the significance of specified features. They demonstrated improved

performance compared to the unweighted method; however, this approach is dependent on the availability of experimental data to identify the significant features.

11. Conclusions

Reduced graphs provide flexible ways of generalising molecular structures while retaining the topology of the original structures. They have proved to be useful for a number of different applications with the optimal graph reduction scheme being dependent on the particular application. For example, for the representation and search of Markush structures, a simple ring/non-ring reduction permits the encoding of generic nomenclature expressions into single nodes, which enables one of the difficulties of handling these structures to be overcome. In applications that aim to identify structure–activity relationship, more complex graph reduction schemes are usually required so that pharmacophoric groups can be identified. It is usually possible to allow the definitions of pharmacophoric features to be determined at run time through the use of SMARTS representations of features such as hydrogen bond donors, hydrogen bond acceptors and ionisable groups. This allows different properties to be emphasised in different applications. Reduced graphs enable similarities to be perceived between heterogeneous compounds, which is beneficial for scaffold-hopping applications and for the capture of SARs from structurally diverse compounds. Furthermore, the small size of the reduced graphs relative to the structures from which they are derived permits the use of graph matching algorithms so that mappings between structures can be generated, which assists in interpreting the results of similarity and SAR analyses.

References

1. Gillet, V. J., Downs, G. M., Ling, A., Lynch, M. F., Venkataram, P., Wood, J. V., and Dethlefsen, W. (1987) Computer-storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical-structure retrieval. *Journal of Chemical Information and Computer Sciences* **27**, 126–137.
2. Rarey, M. and Dixon, J. S. (1998) Feature trees: A new molecular similarity measure based on tree matching. *Journal of Computer-Aided Molecular Design* **12**, 471–490.
3. Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *Journal of Computer-Aided Molecular Design* **15**, 497–520.
4. Stiefl, N., Watson, I. A., Baumann, K., and Zaliani, A. (2006) ErG: 2D pharmacophore descriptions for scaffold hopping. *Journal of Chemical Information and Modeling* **46**, 208–220.
5. Stiefl, N. and Zaliani, A. (2006) A knowledge-based weighting approach to ligand-based virtual screening. *Journal of Chemical Information and Modeling* **46**, 587–596.
6. Gillet, V. J., Downs, G. M., Holliday, J. D., Lynch, M. F., and Dethlefsen, W. (1991) Computer-storage and retrieval of generic chemical structures in patents. 13. Reduced-graph generation. *Journal of Chemical Information and Computer Sciences* **31**, 260–270.

7. Lynch, M. F. and Holliday, J. D. (1996) The Sheffield Generic Structures Project – A retrospective review. *Journal of Chemical Information and Computer Sciences* **36**, 930–936.
8. Shenton, K., Nortin, P., and Fearns, E. A. (1988) Generic Searching of Patent Information, in *Chemical Structures – The International Language of Chemistry* (Warr, W., Ed.), pp 169–178, Springer, Berlin.
9. Fisanick, W. (1990) The chemical abstracts service generic chemical (Markush) structure storage and retrieval capability. Part 1. Basic concepts. *Journal of Chemical Information and Computer Sciences* **30**, 145–154.
10. Ebe, T., Sanderson, K. A. and Wilson, P. S. (1991) The chemical abstracts service generic chemical (Markush) structure storage and retrieval capability. Part 2. The MARPAT file. *Journal of Chemical Information and Computer Sciences* **31**, 31–36.
11. Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom pairs as molecular features in structure activity studies – Definition and applications. *Journal of Chemical Information and Computer Sciences* **25**, 64–73.
12. Willett, P., Winterman, V., and Bawden, D. (1986) Implementation of nearest-neighbor searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences* **26**, 36–41.
13. Brown, N. and Jacoby, E. (2006) On scaffolds and hopping in medicinal chemistry. *MiniReviews in Medicinal Chemistry* **6**, 1217–1229.
14. Daylight. Daylight Chemical Information Systems, Inc., 120 Vantis – Suite 550, Aliso Viejo, CA 92656, USA. www.daylight.com at http://www.daylight.com.
15. Gillet, V. J., Willett, P., and Bradshaw, J. (2003) Similarity searching using reduced graphs. *Journal of Chemical Information and Computer Sciences* **43**, 338–345.
16. Barker, E. J., Gardiner, E. J., Gillet, V. J., Kitts, P., and Morris, J. (2003) Further development of reduced graphs for identifying bioactive compounds. *Journal of Chemical Information and Computer Sciences* **43**, 346–356.
17. Harper, G., Bravi, G. S., Pickett, S. D., Hussain, J., and Green, D. V. S. (2004) The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *Journal of Chemical Information and Computer Sciences* **44**, 2145–2156.
18. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2006) Training similarity measures for specific activities: Application to reduced graphs. *Journal of Chemical Information and Modeling* **46**, 577–586.
19. MDDR. Symyx Technologies Inc, 2440 Camino Ramon, Suite 300, San Ramon, CA 94583. http://www.symyx.com.
20. Takahashi, Y., Sukekawa, M., and Sasaki, S. (1992) Automatic identification of molecular similarity using reduced graph representation of chemical structure. *Journal of Chemical Information and Computer Sciences* **32**, 639–643.
21. Barker, E. J., Cosgrove, D. A., Gardiner, E. J., Gillet, V. J., Kitts, P., and Willett, P. (2006) Scaffold-hopping using clique detection applied to reduced graphs. *Journal of Chemical Information and Modeling* **46**, 503–511.
22. Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **39**, 2887–2893.
23. Gardiner, E. J., Gillet, V. J., Willett, P., and Cosgrove, D. A. (2007) Representing clusters using a maximum common edge substructure algorithm applied to reduced graphs and molecular graphs. *Journal of Chemical Information and Modeling* **47**, 354–366.
24. Bradbury, R. H., Allott, C. P., Dennis, M., Fisher, E., Major, J. S., Masek, B. B., Oldham, A. A., Pearce, R. J., Rankine, N., Revill, J. M., Roberts, D. A., and Russell, S. T. (1992) New nonpeptide angiotensin-II receptor antagonists. 2. Synthesis, biological properties, and structure-activity relationships of 2-alkyl-4-(biphenylmethoxy)quinoline derivatives. *Journal of Medicinal Chemistry* **35**, 4027–4038.
25. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2008) Evolving interpretable structure-activity relationships. 1. Reduced graph queries. *Journal of Chemical Information and Modeling* **48**, 1543–1557.
26. Birchall, K., Gillet, V. J., Harper, G., and Pickett, S. D. (2008) Evolving interpretable structure-activity relationship models. 2. Using multiobjective optimization to derive multiple models. *Journal of Chemical Information and Modeling* **48**, 1558–1570.
27. Birchall, K., Gillet, V. J., Willett, P., Ducrot, P., and Luttmann, C. (2009) Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *Journal of Chemical Information and Modeling* **49**, 1330–1346.
28. Ujvary, I. (1997) BIOSTER: A database of structurally analogous compounds. *Pesticide Science* **51**, 92–95.
29. WOMBAT. Sunset Molecular. Available at http://www.sunsetmolecular.com/.
30. Hessler, G., Zimmermann, M., Matter, H., Evers, A., Naumann, T., Lengauer, T., and Rarey, M. (2005) Multiple-ligand-based virtual screening: Methods and applications of the MTree approach. *Journal of Medicinal Chemistry* **48**, 6575–6584.

Chapter 9

Fragment Descriptors in Structure–Property Modeling and Virtual Screening

Alexandre Varnek

Abstract

This chapter reviews the application of fragment descriptors at different stages of virtual screening: filtering, similarity search, and direct activity assessment using QSAR/QSPR models. Several case studies are considered. It is demonstrated that the power of fragment descriptors stems from their universality, very high computational efficiency, simplicity of interpretation, and versatility.

Key words: Fragmental approach, Fragment descriptors, QSAR, QSPR, Filtering, Similarity, Virtual screening, In silico design

1. Introduction

Chemoinformatics aims to discover active and/or selective ligands for biologically related targets by conducting screening, ideally, of all possible compounds against all possible targets, or at least, in practice, available libraries of compounds against main target families [1]. One can hardly imagine to screen experimentally the chemical universe containing from 10^{12} to 10^{180} drug-like compounds [2] against the biological target universe. Nowadays, the number of experimentally screened compounds does not exceed several millions per biological target, whereas a single inexpensive computational study allows one to screen the libraries up to 10^{12} molecules and this number tends to grow up with the evolution of hardware and related software tools. Therefore, this is not surprising that the virtual, or in silico, screening approaches play a key role in chemogenomics.

Virtual screening is usually defined as a process in which large libraries of compounds are automatically evaluated using computational techniques [3]. Its goal is to discover putative hits in large databases of chemical compounds (usually ligands for biological

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_9, © Springer Science+Business Media, LLC 2011

targets) and remove molecules predicted to be toxic or those possessing unfavorable pharmacodynamic or pharmacokinetic properties. Generally, two types of virtual screening are known: structure-based and ligand-based. The former explicitly uses 3D structure of a biological target at the stage of hit detection, whereas the latter uses only information about structure of small molecules and their properties (activities). Structure-based virtual screening (docking, 3D pharmacophores) has been described in series of review articles, *see* [4–6] and references therein.

In this paper ligand-based virtual screening involving fragment descriptors is discussed. Fragment descriptors represent selected substructures (fragments) of 2D molecular graphs and their occurrences in molecules; they constitute one of the most important types of molecular descriptors [7]. Their main advantage is related to simplicity of their calculation, storage and interpretation (*see* review articles [8–12]). Substructural fragment are information-based descriptors [13] which tend to code the information stored in molecular structures. This contrasts with knowledge-based (or semiempirical) descriptors issued from the mechanistic consideration. Selected descriptors form a "chemical space" in which each molecule is represented as a vector. Due to their versatility, fragment descriptors could be efficiently used to create a chemical space which separates active and non-active compounds.

Historically, molecular fragments were used in first additive schemes developed in the 1950s to estimate physicochemical properties of organic compounds by Tatevskii [14, 15], Bernstein [16], Laidler [17], Benson and Buss [18], and others. The Free–Wilson method [19], one of the first QSAR approaches invented in 1960s, is based on the assumption of the additivity of contributions of structural fragments to the biological activity of the whole molecule. Later on, fragment descriptors were successfully used in expert systems able to classify chemical compounds as active or inactive with respect to certain type of biological activity. Hiller [20, 21], Golender and Rosenblit [22, 23], Piruzyan, Avidon et al. [24, 25], Cramer [26], Brugger, Stuper and Jurs [27, 28], and Hodes et al. [29] pioneered in this field.

An important class of fragmental descriptors, so-called *screens* (structural *keys*, *fingerprints*), has been developed in the 1970s [30–34]. As a rule, they represent bit strings which can effectively be stored and processed by computers. Although their primary role is to provide efficient substructure searching capabilities in large chemical databases, they are efficiently used for similarity searching [35, 36], to cluster large data sets [37, 38], to assess chemical diversity [39], as well as to conduct SAR [40] and QSAR [41] studies. Nowadays, application of modern machine-learning techniques significantly improves predictive performance of structure–property models based on fragment descriptors.

This paper briefly reviews the application of fragment descriptors in virtual screening of large libraries of organic compounds focusing mostly on its three stages: (1) filtering, (2) similarity search, and (3) direct activity/property assessment using QSAR/QSPR models. A particular attention will be paid to new approaches in structure–property modeling (ensemble modeling, applicability domain, inductive learning transfer) and in mining of chemical reactions. Most of examples described here concerns the ISIDA (In SIlico Design and Data Analysis) platform for virtual screening.

2. Types of Fragment Descriptors

Due to their enormous diversity, one could hardly review all types of 2D fragment descriptors used for structural search in chemical database or in SAR/QSAR studies. Here, we focus on some of them which are the most efficiently used in virtual screening and in silico design of organic compounds.

According to Lounkine et al. [42], there exists four major strategies to fragment design: knowledge-based, synthetically oriented, random, and systematic and hierarchical. The knowledge-based methods are based on chemical and pharmaceutical expertise. As examples, one could mention fragments dictionaries for ADME predictions [43] or toxicity alerts [44, 45], and "privileged" substructures recurrent in families of bioactive compounds [46, 47]. In retrosynthetic fragmentation methods, substructures are obtained by breaking bonds in molecules described by cataloged chemical reactions [48, 49]. The main underlying idea is that the resulting fragments can be chemically re-combined in different ways. The most known example of such fragmentation is *Retrosynthetic Combinatorial Analysis Procedure* (RECAP) approach [50] which defines 11 chemical bond types where a cleavage can occur. In random molecular fragmentation methods [42], substructures populations are generated for selected molecules by random deletion of bonds in their connectivity tables followed by sampling of the resulting fragments. Comparing fragments distributions obtained for the set of molecules of given activity class, one can identify class-specific substructures which could be used in virtual screening.

Systematic and hierarchical approaches are based on the predefined rules of fragmentation. Generally, molecular fragments could be classified with respect to their topology (atom-based, chains, cycles, polycycles, etc.), information content of vertices in molecular graphs (atoms, groups of atoms, pharmacophores, descriptor centers) and the level of abstraction when some information concerning atom and bond types is omitted. Some popular fragmentation schemes are discussed below.

Purely structural fragments are used as descriptors in ACD/Labs [51], NASAWIN [52], ISIDA [12], and some other programs. These are 2D subgraphs in which all atoms and/or bonds are represented explicitly and no information about their properties is used. Their typical example is sequences of atoms and/or bonds of variable length, branch fragments, saturated and aromatic cycles (polycycles), and atom-centered fragments (ACF). The latter consist of a single central atom surrounded by one or several shells of atoms with the same topological distance from the central one. The ACF were invented by Tatevskii [14] and Benson and Buss [18] in the 1950s as elements of additive schemes for predicting physicochemical properties of organic compounds. In the early 1970s, Adamson [53] investigated the distribution of one shell ACF in some chemical databases with respect to their possible application as screens. Hodes reinvented one shell ACF as descriptors in SAR studies under the name *augmented atoms* [29], and also suggested *ganglia augmented atoms* [54] representing two shells ACF with generalized second-shell atoms. Later on, one shell ACF were implemented by Baskin et al. in the NASAWIN [52] software and by Solov′ev and Varnek in ISIDA [12] package (*see* Fig. 1). Atom-centered fragments with arbitrary number of shells were implemented by Filimonov and Poroikov in the PASS [55] program as *multilevel neighborhoods of atoms* [56], by Xing and Glen as *tree structured fingerprints* [57], by Bender et al. as *atom environments* [58, 59] and *circular fingerprints* [60–62], and by Faulon as *molecular signatures* [63–65].

It has been found that characterizing atoms only by element types is too specific for similarity searching and therefore does not provide sufficient flexibility needed for large-scaled virtual screening. For that reason, numerous studies were devoted to increase the informational content of fragment descriptors by adding some useful empirical information and/or by representing a part of molecular graph implicitly. The simplest representatives of those descriptors were *atom pairs and topological multiplets* based on the notion of *descriptor center* representing an atom or a group of atoms which could serve as centers of intermolecular interactions. Usually, descriptor centers include heteroatoms,

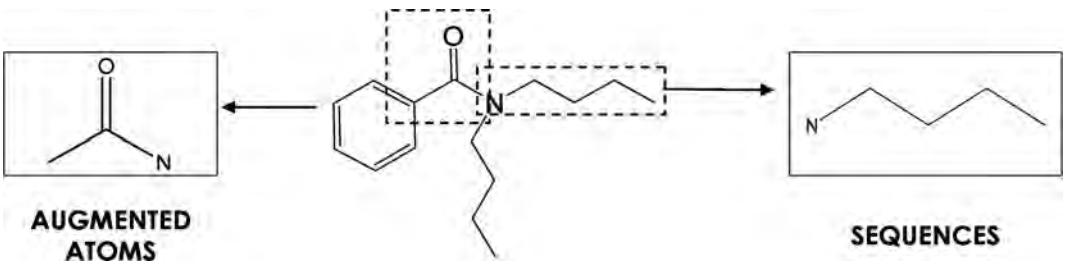


Fig. 1. Decomposition of a chemical structure into fragments. Examples of *sequences* and *augmented atoms* used as descriptors in the ISIDA program [12].

unsaturated bonds and aromatic cycles. An *atom pair* is defined as a pair of atoms (**AT**) or descriptor centers separated by a fixed topological distance: $\mathbf{AT}_i$-$\mathbf{AT}_j$-***Dist***, where $Dist_{ij}$ is the shortest path (the number of bonds) between AT_i and AT_j. Analogously, a *topological multiplet* is defined as a multiplet (usually triplet) of descriptor centers and topological distances between each pair of them. In most of cases, these descriptors are used in binary form in order to indicate the presence or absence of the corresponding features in studied chemical structures.

The atom pairs were first suggested for SAR studies by Avidon under the name *SSFN* (*Substructure Superposition Fragment Notation*) [25, 66]. Then they were independently reinvented by Carhart and co-authors [67] for similarity and trend vector analysis. In contrast to SSFN, Carhart's atom pairs are not necessarily composed only of descriptor centers, but account for the information about element type, the number of bonded non-hydrogen neighbors and the number of π electrons. Nowadays, Carhart's atom pairs are rather popular in virtual screening. *Topological Fuzzy Bipolar Pharmacophore Autocorrelograms* (*TFBPA*) [68] by Horvath are based on atom pairs, in which real atoms are replaced by pharmacophore sites (hydrophobic, aromatic, hydrogen bond acceptor, hydrogen bond donor, cation, anion). These descriptors were successfully applied in virtual screening against a panel of 42 biological targets using similarity search based on several fuzzy and non-fuzzy metrics [69], performing only slightly less well than their 3D counterparts [68]. *Fuzzy Pharmacophore Triplets* (*FPT*) by Horvath [70] is an extention of *TFBPA* for three sites pharmacophores. An important innovation in the *FPT* concerns accounting for proteolytic equilibrium as a function of pH [70]. Due to this feature, even small structural modifications leading to a pK_a shift, may have a profound effect on fuzzy pharmocophore triples. As a result, these descriptors efficiently discriminate structurally similar compounds exhibiting significantly different activities [70] and, therefore, they have been successfully used both in similarity search experiments [68–70] and in structure–property modeling [71].

Some other topological triplets should be mentioned. Thus, *Similog pharmacophoric keys* by Jacoby [72] represent triplets of binary coded types of atoms (pharmacophoric centers) and topological distances between them. Atomic types are generalized by four features (represented as four bits per atom): potential hydrogen bond donor or acceptor; bulkiness, and electropositivity. The *topological pharmacophore-point triangles* implemented in the MOE software [73] represent triplets of MOE atom types separated by binned topological distances. Structure–property models obtained by support vector machine method with these descriptors have been successfully used for virtual screening of COX-2 inhibitors [74] and D_3 dopamine receptor ligands [75].

Topological torsions by Nilakantan et al. [76] is a sequence of four consecutively bonded atoms AT_i-AT_j-AT_k-AT_l, where each atom is characterized by a number of parameters similarly to atoms in Carhart's pairs. In order to enhance efficiency of virtual screening, Kearsley et al. [77] suggested to assign atoms in Carhart's atom pairs and Nilakantan's topological torsions to one of seven classes: cations, anions, neutral hydrogen bond donors, neutral hydrogen bond acceptors, polar atoms, hydrophobic atoms and other. Kuz'min et al. [78, 79] used both connected and disconnected combinations of 4 atoms ("simplex" fragments) as descriptors in SAR/QSAR studies.

In contrast to QSPR studies based mostly on the use of complete (containing all atoms) or hydrogen-suppressed molecular graphs, handling biological activity at the qualitative level, often demands more abstractions. Namely, it is rather convenient to approximate chemical structures by *reduced graphs*, in which each vertex is an atom or a group of atoms (descriptor or pharmacophoric center), whereas each edge is a topological distance $Dist_{ij}$. Such biology-oriented representation of chemical structures was suggested by Avidon et al. as descriptor center connection graphs [25]. Gillet, Willett, and Bradshaw have proposed the GWB-reduced graphs which use the hierarchical organization of vertex labels. This allows one to control the level of their generalization which may explain their high efficiency in similarity searching.

An alternative scheme of reducing molecular graph proposed by Bemis and Murcko [80, 81] involves four-level hierarchical scheme of molecular structure simplification: (1) full molecular structure with all atoms; (2) structure without hydrogen atoms; (3) *scaffolds*, i.e., structures without substituents (which are deleted recursively by means of eliminating the "leaves" of molecular graph); and (4) *molecular frameworks*, i.e., scaffolds, in which all heteroatoms are substituted by carbon atoms, while all multiple bonds are replaced by single bonds. This presentation of molecular graph was found very useful for diversity analysis of large databases [80, 81].

3. Application of Fragment Descriptors in Virtual Screening and In Silico Design

In this chapter, the use of fragment descriptors is considered at different stages of virtual screening: filtering, similarity search, and obtaining and application of SAR/QSAR models.

3.1. Filtering

Filtering is a rule-based approach aimed to perform fast assessment of useful or useless molecules (in the given context). In drug

design, this is used to discard toxic compounds as well as those possessing unfavorable pharmacodynamic or pharmacokinetic properties. Pharmacodynamics considers binding drug-like organic molecules (ligands) to chosen biological target. Since the efficiency of ligand-target interactions depends on spatial complementarity of their binding sites, the filtering is usually performed with 3D-pharmacophores, representing "optimal" spatial arrangements of steric and electronic features of ligands [82, 83]. Pharmacokinetics concerns mostly absorption, distribution, metabolism, and excretion (ADME) related properties: octanol–water partition coefficients (log *P*), solubility in water (log *S*), blood– brain coefficient (log *BB*), partition coefficient between different tissues, skin penetration coefficient, etc.

Fragment descriptors are widely used for early ADME/Tox prediction both explicitly and implicitly. The easiest way to filter large databases concerns detecting undesirable molecular fragments (*structural alerts*). Appropriate lists of structural alerts are published for toxicity [84], mutagenicity [85], and carcinogenicity [86]. Klopman et al. were the first to recognize the potency of using fragmental descriptors for this purpose [87–90]. Their programs CASE [87], MultiCASE [91, 92], as well as more recent MCASE QSAR expert systems [93] proved to be effective tools to assess mutagenicity [88, 92, 93] and carcinogenicity [90, 92] of organic compounds. In these programs, sets of biophores (analogs of structural alerts) were identified and used for activity predictions. A number of more sophisticated fragment-based expert systems of toxicity assessment – DEREK [94], TopKat [95], and Rex [96] – have been developed. DEREK is a knowledge-based system operating with human-coded or automatically generated [97] rules about toxicophores. Fragments in the DEREK knowledge base are defined by means of linear notation language PATRAN which codes the information about atom, bonds and stereochemistry. TopKat uses a large predefined set of fragment descriptors, whereas Rex implements a special kind of atom-pairs descriptors (*links*). To read more information about fragment-based computational assessment of toxicity, including mutagenicity and carcinogenicity, *see* review [98] and references therein.

The most popular filter used in drug design area is based on the Lipinski "rule of five" [99], which takes into account the molecular weight, the number of hydrogen bond donors and acceptors, along with the octanol-water partition coefficient log *P*, to assess the bioavailability of oral drugs. Similar rules of "drug-likeness" or "lead-likeness" were later proposed by by Oprea [100], Veber [101], and Hann [102]. Formally, fragment descriptors are not explicitly involved there. However, many computational approaches to assess log *P* are fragment-based [51, 103, 104]; whereas H-donors and acceptor sites are simplest molecular fragments.

3.2. Similarity Search

The similarity-based virtual screening is based on an assumption that all compounds in a chemical database, which are similar to a query compound, could also have similar biological activities. Although this hypothesis is not always valid (*see* discussion in [105]), quite often the set of retrieved compounds is enriched in actives [106].

To achieve high efficacy of similarity-based screening of databases containing millions compounds, molecular structures are usually represented either by *screens* (structural keys) or by fixed-size or variable-size *fingerprints*. Screens and fingerprints may contain both 2D- and 3D-information. However, the 2D-fingerprints, which are a kind of binary fragment descriptors, dominate in this area. Fragment-based structural keys, like MDL keys [40], are sufficiently good for handling small and medium-sized chemical databases, whereas processing of large databases is performed with fingerprints having much higher information density, such as Daylight [107], BCI [108], and UNITY 2D [109] fingerprints.

The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient T [110]. Two structures are usually considered similar if $T > 0.85$ [106]. Using this threshold and Daylight fingerprints [107], Martin estimated only 30% of a chance to retrieve actives.

In the CATS (*Chemically Advanced Template Search*) approach by Schneider et al. [111], the chemical structures are described by vectors, each component of which is equal to atom pair occurrence divided by the total number of non-hydrogen atoms. Each atom in these atom pairs is attributed to one of five classes: hydrogen bond donor, hydrogen bond acceptor, positively charged, negatively charged, and lipophilic. Topological distances of up to ten bonds are considered in the atom-pair specification. Similarity search with CATS was shown efficient in virtual screening experiments [111].

Hull et al. have developed the *Latent Semantic Structure Indexing* (LaSSI) approach to perform similarity search in low-dimensional chemical space [112, 113]. To reduce the dimension of initial chemical space, the singular value decomposition method is applied for the descriptor-molecule matrix. Ranking molecules by similarity to a query molecule was performed in the reduced space using the cosine similarity measure [114], whereas the Carhart's atom pairs [67] and the Nilakantan's topological torsions [76] were used as descriptors. The authors claim that this approach "has several advantages over analogous ranking in the original descriptor space: matching latent structures is more robust than matching discrete descriptors, choosing the number of singular values provides a rational way to vary the 'fuzziness' of the search" [112].

The issue of "fuzziness" in similarity search was developed by Horvath et al. [68–70] both for the atom pairs and Fuzzy pharmacophore triplets (FPT). The first fuzzy similarity metrics suggested in [68] uses partial similarity scores calculated with respect to the inter-atomic distances distributions for each pharmacophore pair. In that case, the "fuzziness" enables to compare pairs of pharmacophores with different topological or 3D distances. Fuzzy pharmacophore triplets (*see* Subsection 2) can be gradually mapped onto related basis triplets, thus minimizing binary classification artifacts [70]. In the similarity scoring index introduced in reference [70], both simultaneous absence and presence of a pharmacophore triplet in two molecules are taken into account.

Most of similarity search approaches require only a single reference structure. However, in practice several compounds with the same type of biological activity are often available. This motivated Hert et al. [115] to develop the *data fusion method* which allows one to screen a database using all available reference structures. Then, the similarity scores are combined for all retrieved structures using selected fusion rules. Searches conducted on the MDL Drug Data Report database using fragment-based UNITY 2D [109], BCI [108], and Daylight [107] fingerprints have proved the effectiveness of this approach.

The main drawback of the conventional similarity search concerns an inability to use experimental information on biological activity to adjust similarity measures. This leads to inability to discriminate between relevant and non-relevant fragment descriptors being used for computing similarity measures. To tackle this problem, Cramer et al. [26] developed *substructural analysis* in which each fragment (represented as a bit in a fingerprint) is weighted by taking into account its occurrence in active and in inactive compounds. Later on, many similar approaches have been described in the literature [116].

Another way to perform a similarity-based virtual screening is to retrieve the structures containing a user-defined set of "pharmacophoric" features. In the *Dynamic Mapping of Consensus positions* (DMC) algorithm by Godden et al. [117] those features are selected by finding common positions in bit strings for all active compounds. The *potency-scaled DMC* algorithm (POT-DMC) [118] is a modification of DMC in which compounds activities are taken into account. The latter two methods may be considered as intermediate between conventional similarity search and probabilistic SAR approaches.

Batista et al. have developed the MolBlaster method [119], in which molecular similarity is assessed by *Differential Shannon Entropy* [120] computed from populations of randomly generated fragments. For the range $0.64 < T < 0.99$, this similarity measure provides with the same ranking as the Tanimoto index T. However, for the smaller values of T the entropy-based index is

a more sensitive, since it distinguishes between pairs of molecules having almost identical *T*. To adapt this methodology for large-scale virtual screening, the *Proportional Shannon Entropy* (PSE) metrics was introduced [121]. A key feature of this approach is that class-specific PSE of random fragment distributions enables the identification of the molecules sharing a significant number of signature substructures with known active compounds. Another approach based on random fragments has been developed by Lounkine et al. [42]. Comparison of distributions of random fragments obtained for the set of molecules of given activity class, they have extracted *Activity Class Characteristic Substructures* (ACCS) which only occur in compounds having similar activity. Combinations of ACCS carry compound class-specific information and therefore, they can be encoded as class directed fingerprints and used to search databases for novel active compounds.

Another way to perform potency-related similarity search concerns selection of fragment descriptors selected for QSAR modeling. This strategy of "tailored similarity" [122] has been used by Fourches [123] for similarity search of anticonvulsants in Maybridge and NCI databases. At the first step, QSAR models based on fragment descriptors have been obtained for the training set of 48 compounds. Then, atom/bond sequences involved in the model were used to build chemical space in which a similarity search has been performed.

Sometimes, similarity search based on explicit molecular fragments is not able to explain unexpected large activity difference of the molecules, chemical structures of which look similar ("activity cliffs" [124, 125]). In this case, application of topological pharmacophores (especially those accounting for proteolytic equilibrium effects) and chemically meaningful similarity scores could be particularly useful. Figure 2 illustrates a typical strength of

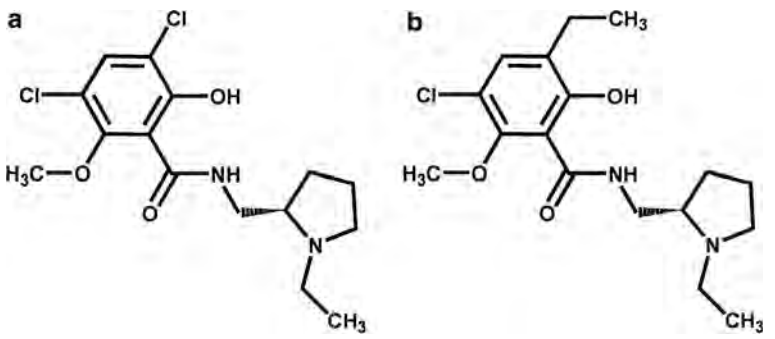


Fig. 2. State-of-the-art similarity evaluations would all agree that the compounds (a) and (b) are virtually identical. However, these molecules actually display significantly different biological activities. Due to its pK_a-sensitive pharmacophore feature flagging scheme, the FPT-based similarity scoring recognizes the difference between them because the compound (a) is an anion, whereas (b) is neutral at pH = 7.

fuzzy pharmacophores driven similarity searching, which is able to explain apparent "activity cliffs". Typically, substitution of an ethyl group by a halogen atom, both flagged as "hydrophobes" by pharmacophore feature flagging routines, would leave the overall pharmacophore pattern unchanged and the two compounds would be virtually undistinguishable (near null dissimilarity score). However, this apparently harmless chemical modification triggers a ionization propensity change of a close proteolytic group and practically toggles the state of an ionic center in the molecule, with important effects on activity. Fuzzy Pharmacophore Triplets [70, 126] successfully take this phenomenon into account and therefore do not overestimate the similarity of the two molecules (Fig. 2).

3.3. SAR/QSAR/QSPR Models

Simplistic and heuristic similarity-based approaches can hardly produce as good predictive models as modern machine-learning methods able to assess biological or physicochemical properties. SAR/QSAR-based virtual screening consists in direct assessment of activity values (qualitative or quantitative) of all compounds in the database followed by selection of hits possessing desirable activity. Generally, approaches of two types – classification and regression – are used in the modeling. The former assesses a probability that a given compound belongs to a particular class (e.g. active or not active) whereas the latter numerically evaluates the activity values. Several examples of SAR/QSAR studies involving fragment descriptors are given below.

Harper et al. [127] have demonstrated a good performance of classification using *binary kernel discrimination* method to screen large databases when Carhart's atom-pairs [67] and Nilakantan's topological torsions [76] are used as descriptors.

Aiming to discover new cognition enhancers, Geronikaki et al. [128] applied the PASS program [55], which implements a probabilistic Bayesian-based approach, and the DEREK rule-based system [94] to screen a database of highly diverse chemical compounds. Eight compounds with the highest probability of cognition-enhancing effect were selected. Experimental tests have shown that all of them possessed a pronounced antiamnesic effect.

Bender et al. [58–62] have applied several classification machine-learning methods (naïve Bayesian classifier, inductive logic programming, and support vector inductive learning programming) in combination with circular fingerprints to perform the classification of bioactive chemical compounds and to carry out virtual screening on several biological targets. It has been shown that the performance of support vector inductive learning programming was significantly better than the other two methods [62].

Regression QSAR/QSPR models are used to assess ADME/Tox properties or to detect "hit" molecules capable to bind a certain biological target. Available in the literature fragments based QSAR models for blood–brain barrier [129], skin permeation rate [130], blood–air [131] and tissue–air partition coefficients [131] could be mentioned as examples. Many theoretical approaches of calculation of octanol-water partition coefficient log P involve fragment descriptors. The methods by Rekker [132, 133], Leo and Hansch (CLOGP) [103, 134], Ghose-Crippen (ALOGP) [135–137], Wildman and Crippen [138], Suzuki and Kudo (CHEMICALC-2) [139], Convard (SMILOGP) [140], and Wang (XLOGP) [141, 142] represent just a few modern examples. Fragment-based predictive models for estimation solubility in water [143] and DMSO [143] are available.

Benchmarking studies performed in references [129–131, 144] show that QSAR/QSPR models for various biological and physicochemical properties involving fragment descriptors are, at least, as robust as those involving topological, quantum, electrostatic and other types of descriptors.

In fact, classical QSAR has been developed for relatively small congener datasets. Below, we describe some strategies to improve predictive performance of the models developed on relatively large or too small structurally diverse datasets: ensemble modeling, "divide and conquer" technique and inductive learning transfer approach implemented into the ISIDA program package. It should be noted that some ISIDA models for biological activities, ADME properties, aqueous solubility, and stability constants of metals in solution are available for the users via INTERNET interface at http://infochim.u-strasbg.fr.

3.3.1. Ensemble Modeling

Relationships between chemical structures of compounds and their properties may have a very complex nature. As a consequence, a single QSAR approach may be insufficient to reflect the structure–activity relationships accurately enough. Therefore, in order to improve performance of predictions, one could use Consensus Model (CM) approach which combines several individual models [144–146]. There exist several possible ways to generate ensembles of models. The ISIDA program builds many individual models on one same training set issued from different initial pools of descriptors, each of which corresponds to a given fragmentation type. Only models for which leave-one out cross-validation correlation coefficient Q^2 is larger than a user-defined threshold are selected. Then, for each query compound, the program calculates the predicted property as an arithmetic mean of values obtained with the selected models (excluding, according to Grubbs's test [147], those leading to outlying values). This approach has been successfully used to obtain predictive MLR models for various ADME related properties (Skin Permeation

Rate [129], Blood–Air and Tissue–Air Partition Coefficients [131], Blood–Brain Barrier Permeation [129]), some biological activities [148], thermodynamic parameters of metal complexation and extraction [149, 150], free energies of hydrogen-bond complexes [12], and melting points of ionic liquids [144]. The Stochastic QSAR Sampler (SQS) program builds individual models issued from different descriptor subsets selected in genetic algorithm-based variable selection procedure. SQS has been successfully used to build predictive consensus models involving FPT descriptors for various biological activities [71, 151].

One can also build consensus model combining different machine-learning approaches. Recently, QSAR models for aquatic toxicity ($pIGC_{50}$) of organic molecules against *Tetrahymena pyriformis* have been obtained in the framework of collaborative project between six research teams [152]. Initial dataset was randomly split into a training set (644 compounds) and a test set (339 compounds) to afford an external validation of training set models. Incidentally, a second test set (110 compounds) has become available after the model building was completed. Each group used their favorite machine-learning approaches and descriptors (Dragon [153], MolconnZ [153] and ISIDA fragments), as well as their definition of models applicability domains. Totally, from 9 to 15 individual models have been obtained and applied to predict $pIGC_{50}$ for compounds in both test sets. Applicability domain assessment leads to significant improvement of the prediction accuracy for both test sets but dramatically reduces the chemical space coverage of the models. To increase the model coverage, different types of consensus models were developed by averaging predicted toxicity values computed from all models, with or without taking into account their respective applicability domains. In all cases, consensus models leads to better prediction accuracy for the external test sets as well as to the largest space coverage, as compared to any individual constituent model. Thus, on the Regression Error Curves plot, the curve corresponding to the consensus model lays higher than the curves of individual models (Fig. 3).

It has been shown [151, 154] that the variance of predicted value for a query molecule could be used as criterion of the prediction performance of the consensus models. More individual models converge toward one number, more reliable the predicted value is.

3.3.2. "Divide and Conquer" Technique

For large structurally diverse datasets, "*Divide and Conquer*" (DC) strategy could be particularly useful. It consists in split of the initial dataset into smaller congener subsets followed by obtaining the *local* QSAR models on each subset. The local models together with *global* ones obtained for the whole initial set are then used for consensus model calculations on the external test

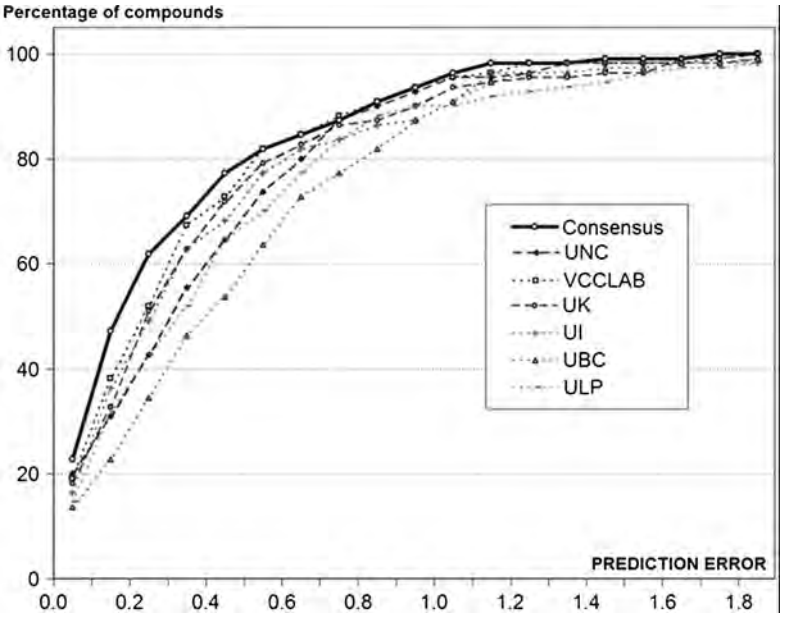


Fig. 3. Percentage of compounds for the test set two (containing 110 compounds) versus prediction errors [152]. Individual models were prepared by six teams participating in the project: *UNC* University of North Carolina at Chapel Hill in USA, *ULP* Louis Pasteur University in France, *UI* University of Insubria in Italy, *UK* University of Kalmar in Sweden, *VCCLAB* Virtual Computational Chemistry Laboratory in Germany, and *UBC* University of British Columbia in Canada.

set. In that case, the applicability domain must be respected in order to avoid application the model to the query compounds dissimilar to the related training (sub)set.

The DC approach [145] has been applied to develop QSAR models for intrinsic aqueous solubility (log *S*) for the set of 1,630 compounds compiled from the references [155–158]. The initial set has been split onto four subsets using an algorithm combing both hierarchical and non-hierarchical clustering approaches [12, 159]. Both for the initial set and for each of subsets, several individual linear models have been selected according to leave-one out cross-validation correlation coefficient. The prediction performance of the models has been tested on the external test set of 412 compounds. Two consensus models (CM) were used: conventional CM involving only global models, and DC–CM involving both global and local models. For the linear correlation log *S* (predicted) versus. log *S* (experimental), root-mean squared error of DC–CM (0.86 log *S* units) is significantly smaller than that of the conventional CM (1.13 log *S* units). Thus, these calculations demonstrate that predictive performance of the DC models significantly outperforms that of linear conventional models [145].

3.3.3. Inductive Learning Transfer Approach

QSAR modeling of pharmacokinetics properties represent of real challenge because in many cases experimental data are available only for relatively small and structurally diverse data sets. In conventional calculations (Single Task Learning, STL), the models are developed for a given property "from the scratch" without any involvements of available information for other related properties. For small initial data sets, they may fail because of lack of experimental information. In that case, Multi-Task Learning (MTL) and Feature Net (FN) [160] approaches integrating the knowledge extracted from different data sets could become a reasonable solution. In MTL, the knowledge is cumulated when the models are simultaneously trained for several related properties. In FN, estimated values of related properties are used as descriptors.

Higher performance of MTL and FN approaches over STL has been demonstrated in QSAR modeling of tissue–air partition coefficients (log K) [145, 161]. The initial dataset contained 11 different log K types obtained for a diverse set containing 199 organic compounds for human (H) and rat (R) species [131]. For each molecule, experimental data were not systematically available for all tissues and species. Thus, individual datasets included, respectively, 138, 35, 42, 30, 34 and 38 compounds for H-blood, H-brain, H-fat, H-liver, H-kidney and H-muscle, and 59, 99, 100, 27 and 97 compounds for R-brain, R-fat, R-liver, R-kidney and R-muscle partition coefficients. Associative Neural Networks (ASNN) approach [162] and fragment descriptors were used to build the models. In three layers neural networks, each neuron in the initial layer corresponded to one molecular descriptor. Hidden layer contained from three to six neurons, whereas the output layer contained one (for STL and FN) or 11 (MTL) neurons, corresponding to the number of simultaneously treated properties. In STL and MTL calculations, only fragment descriptors were used as an input. In FN calculations, the models were built only for one target property, whereas other ten properties served as complementary descriptors. Each model was validated using external fivefold cross-validation procedure [163]. The model was accepted if squared determination coefficient (R^2) for the linear correlation between predicted and experimental property values exceeds a threshold of $R^2 > 0.5$. Figure 4 shows that conventional STL modeling results to predictive models only for four properties corresponding to relatively large (about 100 compounds and more) data sets: H-blood, R-fat, R-liver, and R-muscle. Application of MTL and FN approaches allowed us to significantly improve the reliability of the calculations: predictive models were obtained for nine types of partition coefficients tissue–air (Fig. 4), *see* details in reference [161].

3.3.4. Applicability Domain

The question arises whether QSAR models built on the training set of limited size (usually, several hundred molecules) could be

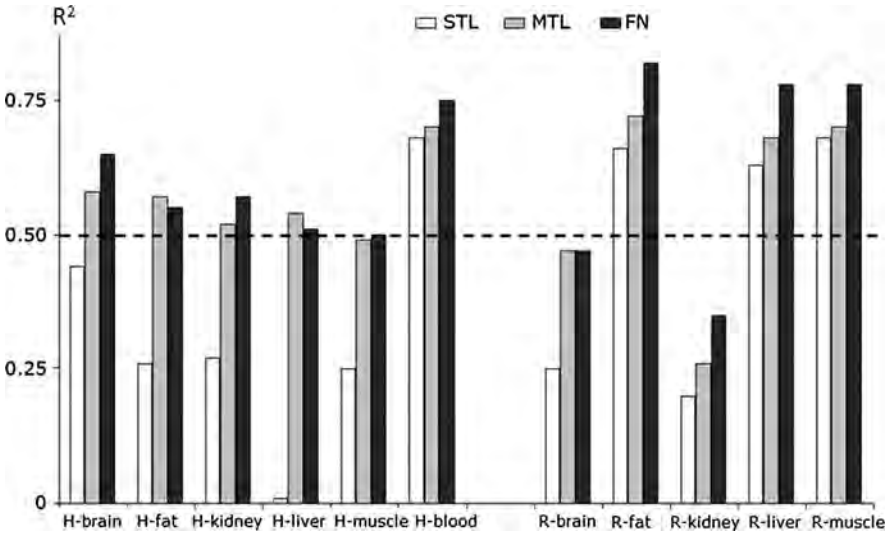


Fig. 4. Performance of different learning strategies to predict Human or Rat air tissue partition coefficient. MTL and FN calculations involved all 11 studied properties. The horizontal line at $R^2 > 0.5$ corresponds to model acceptance threshold (*see* details in [161]).

successfully applied to predict properties of the molecules different from training examples? The question is not trivial because defining an applicability domain (AD) amounts to the calibration of a meta-model based on its own specific attributes and equations. For a given query molecule, the AD is supposed to assess a "predictability score" allowing user to take a decision concerning the application of QSAR model associated to this AD.

AD definition research is nowadays a hot topic in the QSAR field. Typically, state-of-the-art AD models can be roughly classified into range-based [164–166], distance-based [167–169] and density-based [167–171] approaches which could be applied to the models involving any type of descriptors. On the other hand, there exist some AD approaches specifically related to fragment descriptors [145, 149, 172]. Thus, the *Fragment Control* (FC) algorithm [145, 149, 172] prevents to apply a given model to query molecule containing molecular fragments which do not occur in the initial pool of descriptors. This is a very simple but rather efficient approach which may significantly improve the quality of predictions. Figure 5 shows that statistical parameters of the models accepted by FC are rather close to those obtained in external cross-validation for the training set, which is not the case for the rejected models.

Unlike FC, *Model Fragment Control* (MFC) technique [172] is model-dependent and deals with only fragment descriptors involved in the model. Let us suggest that a given model involves N_{tot} descriptors. Each individual molecule in the training set contains $N_i \leq N_{\text{tot}}$ descriptors, where N_i varies from $N_{\min}$ to

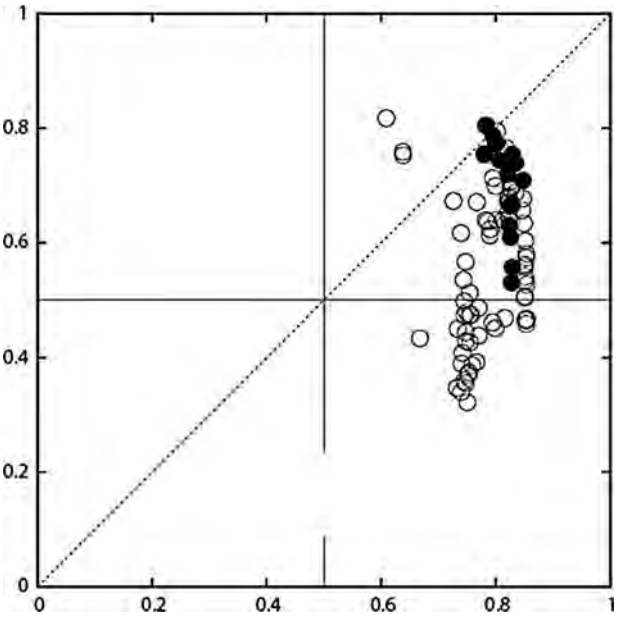


Fig. 5. Predictive performance of individual classification models developed for antibiotic activity for the compounds from the French National Library with Naïve Bayes method. ROC AUC of cross-validation for the training set (4,563 compounds, 62 actives) versus that obtained for the external test set (122 molecules). Each point represents a model issued from different initial descriptors pool with (*black cycles*) or without (*empty cycles*) accounting for the model's applicability domain. The *horizontal* and *vertical lines* at ROC AUC = 0.5 correspond to the prediction performances of a random selection ("no model").

N_{max} (N_{min}, $N_{max} \leq N_{tot}$). The query compound is discarded if the related N_{query} value is outside of this range. MCF is a robust procedure for discarding meaningless queries [172]. Thus, MCF discarded alkans (which are definitely not metal binders) from the set of molecules where QSPR models for metal complexation have been applied. Surprisingly, the range-based, distance-based and FC methods do not recognize alkans as being outside of AD.

Applying an AD, one should always look for a trade-off between accuracy of prediction and test set coverage – the number of the query molecules accepted by AD. More restrictive AD as a better prediction performance can be achieved for smaller fraction of the test set. Application of consensus models may significantly increase the test set coverage. The point is that AD discards some individual models involved in the consensus model, but usually not all of them. This has been demonstrated by Zhu et al. [152, 154] in the combinatorial QSAR modeling of chemical toxicants tested against *T. pyriformis*. The consensus model involving nine individual models was applied to two external test sets. Each individual model was associated with its applicability domain. The Mean Average Error (MAE) of predictions with the

consensus model (0.27 and 0.34 $pIGC_{50}$ units for test sets one and two, respectively) was equal or lower than that for the individual models (0.27–0.44 and 0.33–0.43, respectively). On the other hand, the consensus calculation covered all molecules in the test sets one and two, whereas the coverage of the individual models varied from 80 to 97%, and from 43 to 98%; respectively. Thus, a joint application of the ensemble modeling and applicability domain approaches leads to reasonable balance between test set coverage and prediction accuracy [152, 154].

3.4. In Silico Design

In this section, we consider examples of virtual screening performed on a database containing only virtual (still non-synthesized or unavailable) compounds. Generation of virtual libraries is usually performed using combinatorial chemistry approaches [173–175]. One of simplest ways is to attach systematically user-defined substituents R_1, R_2, …, R_N to a given scaffold. If the list for the substituent R_i contains n_i candidates, the total number of generated structures is $N = \prod_i n_i$, although taking symmetry into account could reduce the library's size. The number n_i of substituents R_i should be carefully selected in order to avoid a generation of too large set of structures (combinatorial explosion). The "optimal" substituents could be prepared using fragments selected at the QSAR stage, since their contributions into activity (for linear models) allow one to estimate an impact of combining the fragment into larger species (R_i). In such a way, a focused combinatorial library could be generated.

The technology based on combining QSAR, generation of virtual libraries and screening stages has been implemented into ISIDA and applied to computer-aided design of new uranyl binders belonging to two different families of organic molecules: phosphoryl containing podands [176] and monoamides [146]. QSAR models have been developed using different machine-learning methods (multi-linear regression analysis, associative neural networks [177] and support vector machines [178]) and fragment descriptors (atom/bond sequences and augmented atoms). Then, these models were used to screen virtual combinatorial libraries containing up to 11,000 compounds. Selected hits were synthesized and tested experimentally. Experimental data correspond well to predicted uranyl binding affinity. Thus, initial data sets were significantly enriched with new efficient uranyl binders, and one of hits was found more efficient than previously studied compounds. A similar study was conducted for development of new 1-[2-(hydroxyethoxy)methyl]-6-(phenylthio) thymine (HEPT) derivatives potentially possessing high anti-HIV activity [148]. This demonstrates universality of fragment descriptors and broad perspectives of their use in virtual screening and in silico design.

4. Mining Chemical Reactions Data Using Condensed Reaction Graphs Approach

Compared to the huge number of reported QSAR and similarity search applications to datasets of individual molecules, very few articles are devoted to chemical reactions. Indeed, chemical reactions are difficult objects because they involve several species of two different types: reactants and products. The "Condensed Graph of Reaction" (CGR) approach [179–181] opens new perspectives in the mining of reaction databases since it allows one to transform several 2D molecular graphs describing a chemical reaction into one single graph. Besides conventional chemical bonds (simple, double, aromatic, etc.), a CGR contains dynamical bonds corresponding to created, broken or transformed bonds. Thus, a chemical reactions database can be transformed into a set of "pseudo-compounds" to which most chemoinformatics methods developed for individual molecules can be applied. Here, we briefly discuss application of CGR approach for the reactions classification, similarity search and quantitative structure–reactivity modeling.

Reactions Classification. A possibility to use CGRs for the analysis of the content of reaction databases has been described in [182]. A sample containing 3,983 Diels-Alder (DA) and 736 metathesis (MT) reactions has been selected from the *ChemInform* and *Reflib* databases using queries given on Fig. 6. All selected reactions were then transformed into CGRs followed by their fragmentation into atom/bond sequences containing at least one dynamical bond. Then, the hierarchical clustering has been performed using Tanimoto similarity coefficient as a metrics. This resulted in four distinct clusters two of which contained exclusively MT reactions, one cluster included exclusively DA reactions,

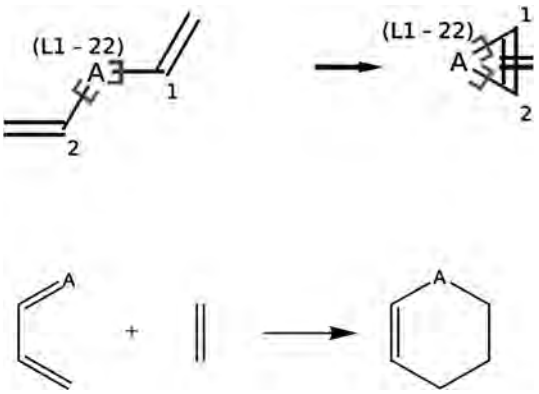


Fig. 6. Queries used for substructural search of metathesis (*top*) and Diels–Alder (*bottom*) reactions in the *ChemInform* and *Reflib* databases. "A" corresponds to any non-hydrogen atom.

and one cluster contained a mixture of DA and MT reactions. Detailed analysis of this mixed cluster shows that 28 reactions initially attributed to MT, in fact, represent *Domino Heck-Diels-Alder* (DHDA) reactions proceeding in two steps: a single bond formation between carbon atoms 1 and 2 followed by the cyclization step according to DA mechanism. Analysis of the clusters contents resulted in preparation of "reaction signatures" for each type of reactions (Fig. 7). One may see that the CGRs of DHDA and DA are very similar: the only difference concerns one dynamical bond which is "created double" for DHDA and "created single" for DA. Substructural search using CGRs on Fig. 7 as queries perfectly separates MT and DHDA reactions which is not always possible using canonical representation of both reactions themselves and reactions queries.

Reaction Similarity Search. As any molecular graphs, CGRs can be fragmented and related fingerprints could be then used for the similarity search. The pertinence of this search depends on the fragments type. Thus, a similarity search using as query the CGR for the domino Heck–Diels–Alder reaction on Fig. 8 and Tanimoto coefficient (T) as metrics has been performed using (1) fragments containing, at least, one dynamical bond, and, (2) fragments containing only dynamical bond(s) conjugated with canonical double bonds. In search (1), for $T > 0.8$, only 9 of 28 DHDA reactions have been retrieved, whereas the all 28 reactions have been found in the search (2).

Structure–Reactivity Relationships. Conventional QSAR modeling of the thermodynamic, kinetic or any other parameters of chemical reactions involving many species is a big problem because

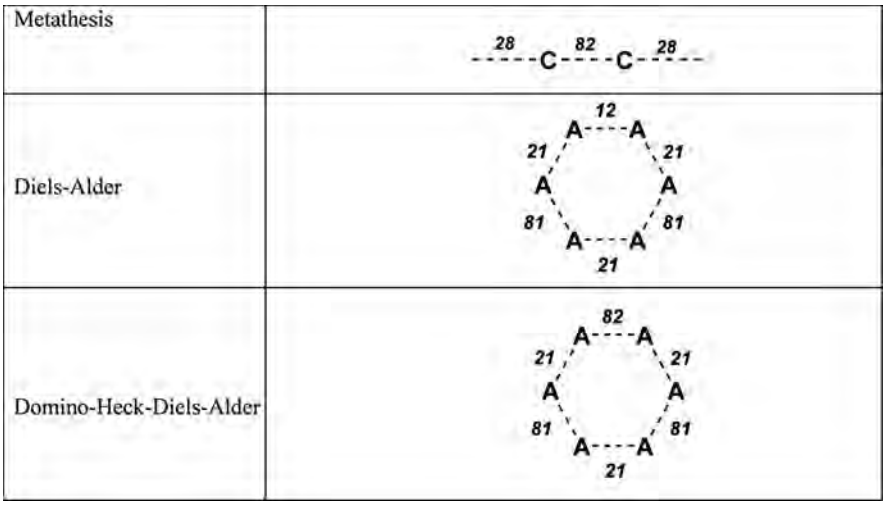


Fig. 7. Examples of CGRs recommended as queries for substructural search for metathesis, Diels–Alder, and Domino Heck–Diels–Alder reactions. The numbers correspond to the types of dynamical bonds (*see* Fig. 9.8). Type "12" corresponds to single bond transformed to double bond.

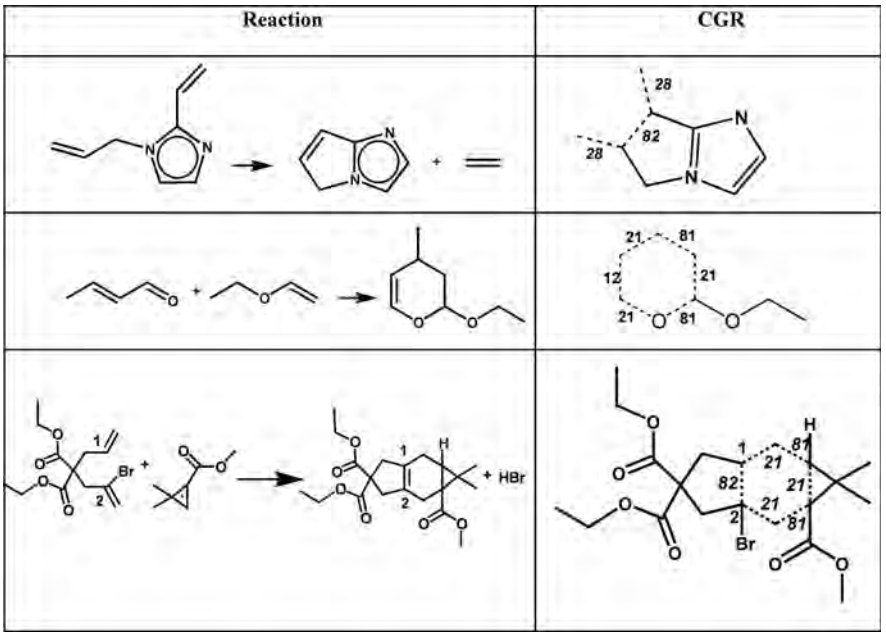


Fig. 8. Typical examples of metathesis (*top*), Diels–Alder (*middle*) and Domino Heck–Diels–Alder (*bottom*) reactions and related Condensed Reaction Graphs (CGR). The following labels are used for the dynamical bonds in GCR: 21 (double bond transformed to single bond), 28 (broken double bond), 82 (created double bond), 81 (created single bond).

it is not clear for which species the descriptors should be calculated. In this situation, CGR could become a reasonable solution since it represents a simple chemical graph for which fragment descriptors can be easily calculated. Recently [183], the CGR approach has been used to build predictive models for reaction rate (log k). The training set contained 463 structurally diverse S_N2 reactions in water at different temperature (totally 1,014 data). ISIDA fragments and reverse temperature have been used as descriptors in SVM calculations. The models have been validated in external tenfold cross-validation procedure repeated ten times after randomization of the data set. Figure 9 shows that log k is reasonably well predicted: squared determination coefficient $R^2 = 0.6$ and root-mean squared error is 1.14. The latter is rather close to the experimental error estimated as 1 log k unit.

5. Limitations of Fragment Descriptors

Despite many advantages of fragment descriptors, they are not devoid of certain drawbacks, which deserve serious attention. Two main problems should be mentioned: (1) "missing fragments" [184] and (2) modeling of stereochemically dependent properties.

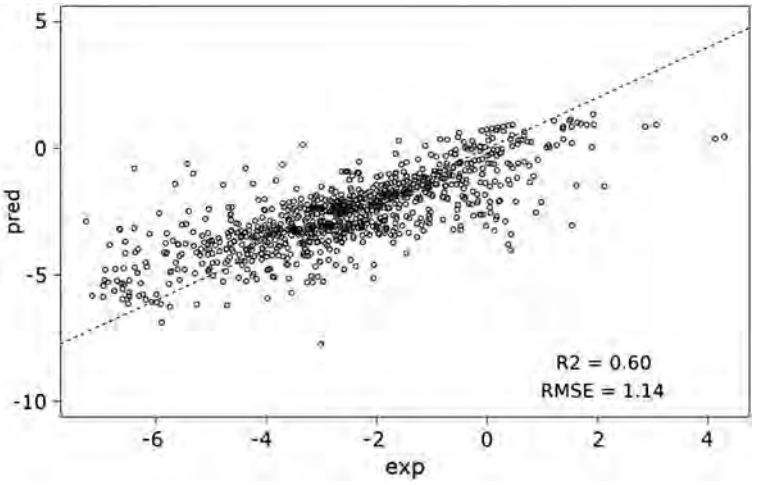


Fig. 9. Predicted versus experimental values of rate constant of S_N2 reactions in water (log *k*). The models were built on the set of 1,014 reactions using SVM method and fragment descriptors extracted from Condensed Graphs of Reactions.

The term "missing fragments" concerns comparison of the lists of fragments generated for the training and test sets. Test set molecules may contain fragments different from those in the initial pool calculated for the training set. The question arises whether the model built from that initial pool can be applied to those test set molecules? This is a difficult problem because a priori it is not clear if the "missing fragments" are important for the property being predicted. Several possible strategies to treat this problem have been reported. The ALOGPS program [162] predicting lipophilicity and aqueous solubility of chemical compounds, flags calculations as unreliable if the analyzed molecule contains one or more E-state atom or bond types missed in the training set. In such a way, the program detects about 90% of large prediction errors [184]. The ISIDA program [12] applies Fragment Control and Model Fragment Control AD to consensus models, which improve the accuracy of prediction at reasonably high coverage of test sets. The NASAWIN program [185], for each model, creates a list of "important" fragments including cycles and all one atom fragments. The test molecule is rejected if its list of "important" fragments contains those absent in the training set [186]. The LOGP program for lipophilicity predictions [187] uses a set of empirical rules to calculate the contribution of missed fragments.

The second problem of using fragment descriptors deals with accounting for stereochemical information. In fact, its adequate treatment is not possible at the graph-theoretical level and requires explicit consideration of hypergraphs. However, in practice, it is sufficient to introduce special labels indicating

stereochemical configuration of chiral centers or E/Z isomers around double bond, then to use them in specification of molecular fragments. Such approach has been used in hologram fragment descriptors [188] as well as in the PARTAN language [97].

6. Conclusion

Fragment descriptors constitute one of the most universal type of molecular descriptors. The scope of their application encompasses almost all existing areas of SAR/QSAR/QSPR studies. Their universality stems from the basic character of the structural theory in chemistry as well as from the fundamental possibility of molecular graph invariants to be expressed in terms of subgraph occurrence numbers [189]. The main advantages of fragment descriptors lie in the ease of their computation as well as in the natural character of structural interpretation of SAR/QSAR/QSPR models. Due to all these factors, fragment descriptors play very important role in structure–property studies and ligand-based virtual screening.

Acknowledgment

The author thanks I. Baskin, D. Horvath, N. Lachiche, F. Hoonakker, A. Wagner, O. Klimchuk and G. Marcou for fruitful discussion and help for preparation of the manuscript.

References

1. Kubinyi, H., and Muler, G. (2004) *Chemogenomics in Drug Discovery*, Wiley-VCH Publishers, Weinheim.
2. Gorse, A. D. (2006) Diversity in Medicinal Chemistry Space. *Curr. Top. Med. Chem.* **6**, 3–18.
3. Walters, W. P., Stahl, M. T., and Murcko, M. A. (1998) Virtual Screening – An Overview. *Drug Discov. Today* **3**, 160–178.
4. Seifert, M. H., Kraus, J., and Kramer, B. (2007) Virtual High-Throughput Screening of Molecular Databases. *Curr. Opin. Drug. Discov. Dev.* **10**, 298–307.
5. Cavasotto, C. N., and Orry, A. J. (2007) Ligand Docking and Structure-Based Virtual Screening in Drug Discovery. *Curr. Top. Med. Chem.* **7**, 1006–1014.
6. Ghosh, S., Nie, A., An, J., and Huang, Z. (2006) Structure-Based Virtual Screening of Chemical Libraries for Drug Discovery. *Curr. Opin. Chem. Biol.* **10**, 194–202.
7. Todeschini, R., and Consonni, V. (2000) *Handbook of Molecular Descriptors*. Wiley-VCH Publishers, Weinheim.
8. Zefirov, N. S., and Palyulin, V. A. (2002) Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci.* **42**, 1112–1122.
9. Japertas, P., Didziapetris, R., and Petrauskas, A. (2002) Fragmental Methods in the Design of New Compounds. Applications of The Advanced Algorithm Builder. *Quant. Struct. Act. Relat.* **21**, 23–37.
10. Artemenko, N. V., Baskin, I. I., Palyulin, V. A., and Zefirov, N. S. (2003) Artificial Neural Network and Fragmental Approach in Prediction of Physicochemical Properties of Organic Compounds. *Russ. Chem. Bull.* **52**, 20–29.

11. Merlot, C., Domine, D., and Church, D. J. (2002) Fragment Analysis in Small Molecule Discovery. *Curr. Opin. Drug Discov. Dev.* **5**, 391–399.
12. Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. P. (2005) Substructural Fragments: An Universal Language to Encode Reactions, Molecular and Supramolecular Structures. *J. Comput. Aided Mol. Des.* **19**, 693–703.
13. Jelfs, S., Ertl, P., and Selzer, P. (2007) Estimation of pKa for Drug Like Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **47**, 450–459.
14. Tatevskii, V. M. (1950) Chemical Structure of Hydrocarbons and Their Heats of Formation. *Dokl. Akad. Nauk SSSR* **75**, 819–822.
15. Tatevskii, V. M., Mendzheritskii, E. A., and Korobov, V. (1951) The Additive Scheme of the Heat of Formation of Hydrocarbons and the Problem of the Heat of Sublimation of Graphite. *Vestn. Mosk. Univ.* **6**, 83–86.
16. Bernstein, H. J. (1952) The Physical Properties of Molecules in Relation to Their Structure. I: Relations Between Additive Molecular Properties in Several Homologous Series. *J. Chem. Phys.* **20**, 263–269.
17. Laidler, K. J. (1956) System of Molecular Thermochemistry for Organic Gases and Liquids. *Can. J. Chem.* **34**, 626–648.
18. Benson, S. W., and Buss, J. H. (1958) Additivity Rules for the Estimation of Molecular Properties: Thermodynamic Properties. *J. Chem. Phys.* **29**, 546–572.
19. Free, S. M., Jr., and Wilson, J. W. (1964) A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **7**, 395–399.
20. Hiller, S. A., Golender, V. E., Rosenblit, A. B., Rastrigin, L. A., and Glaz, A. B. (1973) Cybernetic Methods of Drug Design. I: Statement of the Problem – The Perceptron Approach. *Comput. Biomed. Res.* **6**, 411–421.
21. Hiller, S. A., Glaz, A. B., Rastrigin, L. A., and Rosenblit, A. B. (1971) Recognition of Phisiological Activity of Chemical Compounds on Perceptron with Random Adaptation of Structure. *Dokl. Akad. Nauk SSSR* **199**, 851–853.
22. Golender, V. E., and Rozenblit, A. B. (1974) Interactive System for Recognition of Biological Activity Features in Complex Chemical Compounds. *Avtomatika i Telemekhanika* 99–105.
23. Golender, V. E., and Rozenblit, A. B. (1980) Logico-Structural Approach to Computer-Assisted Drug Design. *Med. Chem.* **11**, 299–337.
24. Piruzyan, L. A., Avidon, V. V., Rozenblit, A. B., Arolovich, V. S., Golender, V. E., Kozlova, S. P., Mikhailovskii, E. M., and Gavrishchuk, E. G. (1977) Statistical Study of an Information File on Biologically Active Compounds: Data Bank of the Structure and Activity of Chemical Compounds. *Khimiko-Farmatsevticheskii Zhurnal* **11**, 35–40.
25. Avidon, V. V., Pomerantsev, I. A., Golender, V. E., and Rozenblit, A. B. (1982) Structure-Activity Relationship Oriented Languages for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **22**, 207–214.
26. Cramer, R. D., III, Redl, G., and Berkoff, C. E. (1974) Substructural analysis: A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **17**, 533–535.
27. Stuper, A. J., and Jurs, P. C. (1976) ADAPT: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques. *J. Chem. Inf. Model.* **16**, 99–105.
28. Brugger, W. E., Stuper, A. J., and Jurs, P. C. (1976) Generation of Descriptors from Molecular Structures. *J. Chem. Inf. Model.* **16**, 105–110.
29. Hodes, L., Hazard, G. F., Geran, R. I., and Richman, S. (1977) A Statistical-Heuristic Methods for Automated Selection of Drugs for Screening. *J. Med. Chem.* **20**, 469–475.
30. Milne, M., Lefkovitz, D., Hill, H., and Powers, R. (1972) Search of CA Registry (1.25 Million Compounds) with the Topological Screens System. *J. Chem. Doc.* **12**, 183–189.
31. Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M. (1973) Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **13**, 153–157.
32. Feldman, A., and Hodes, L. (1975) An Efficient Design for Chemical Structure Searching. I: The Screens. *J. Chem. Inf. Model.* **15**, 147–152.
33. Willett, P. (1979) A Screen Set Generation Algorithm. *J. Chem. Inf. Model.* **19**, 159–162.
34. Willett, P. (1979) The Effect of Screen Set Size on Retrieval from Chemical Substructure Search Systems. *J. Chem. Inf. Model.* **19**, 253–255.
35. Willett, P., Winterman, V., and Bawden, D. (1986) Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System. *J. Chem. Inf. Model.* **26**, 36–41.
36. Fisanick, W., Lipkus, A. H., and Rusinko, A. (1994) Similarity Searching on CAS Registry Substances. 2: 2D Structural Similarity. *J. Chem. Inf. Model.* **34**, 130–140.

37. Hodes, L. (1989) Clustering a Large Number of Compounds. 1: Establishing the Method on an Initial Sample. *J. Chem. Inf. Model.* **29**, 66–71.
38. McGregor, M. J., and Pallai, P. V. (1997) Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Model.* **37**, 443–448.
39. Turner, D. B., Tyrrell, S. M., and Willett, P. (1997) Rapid Quantification of Molecular Diversity for Selective Database Acquisition. *J. Chem. Inf. Model.* **37**, 18–22.
40. Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002) Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **42**, 1273–1280.
41. Tong, W., Lowis, D. R., Perkins, R., Chen, Y., Welsh, W. J., Goddette, D. W., Heritage, T. W., and Sheehan, D. M. (1998) Evaluation of Quantitative Structure-Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor. *J. Chem. Inf. Model.* **38**, 669–677.
42. Lounkine, E., Batista, J., and Bajorath, J. (2008) Random Molecular Fragment Methods in Computational Medicinal Chemistry. *Curr. Med. Chem.* **15**, 2108–2121.
43. Clark, M. (2005) Generalized Fragment-Substructure Based Property Prediction Method. *J. Chem. Inf. Model.* **45**, 30–38.
44. Matter, H., Baringhaus, K. H., Naumann, T., Klabunde, T., and Pirard, B. (2001) Computational Approaches Towards the Rational Design of Drug-Like Compound Libraries. *Comb. Chem. High Throughput Screen.* **4**, 453–475.
45. Oprea, T., Davis, A., Teague, S., and Leeson, P. (2001) Is There a Difference Between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **41**, 1308–1315.
46. Patchett, A. A. N., and Nargund, R. P. (2000) Privileged Structures: An Update. *Annu. Rep. Med. Chem.* **35**, 289–298.
47. Aronov, A. M., McClain, B., Moody, C. S., and Murcko, M. A. (2008) Kinase-Likeness and Kinase-Priviledged Fragments: Toward Virtual Pharmacology. *J. Med. Chem.* **51**, 1214–1222.
48. Gillet, V. M., Myatt G., Zsoldos, Z., and Johnson, P. (1995) SPROUT, HIPPO and CAESA: Tools for De Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discov. Des.* **3**, 34–50.
49. Schneider, G. F., and Fechner, U. (2005) Computer-Based De Novo Design of Drug-Like Molecules. *Nat. Rev. Drug. Discov.* **4**, 649–663.
50. Lewell, X. Q., Judd D. B., Watson, S. P., and Hann, M. M. (1998) RECAP – Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **38**, 511–522.
51. Petrauskas, A. A., and Kolovanov, E. A. (2000) ACD/Log P Method Description. *Perspect. Drug Discov. Des.* **19**, 99–116.
52. Artemenko, N. V., Baskin, I. I., Palyulin, V. A., and Zefirov, N. S. (2001) Prediction of Physical Properties of Organic Compounds Using Artificial Neural Networks Within the Substructure Approach. *Dokl. Chem.* **381**, 317–320.
53. Adamson, G. W., Lynch, M. F., and Town, W. G. (1971) Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II: Atom-Centered Fragments. *J. Chem. Soc. C*, 3702–3706.
54. Hodes, L. (1981) Selection of Molecular Fragment Features for Structure-Activity Studies in Antitumor Screening. *J. Chem. Inf. Comput. Sci.* **21**, 132–136.
55. Poroikov, V. V., Filimonov, D. A., Borodina, Y. V., Lagunin, A. A., and Kos, A. (2000) Robustness of Biological Activity Spectra Predicting by Computer Program Pass for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **40**, 1349–1355.
56. Filimonov, D., Poroikov, V., Borodina, Y., and Gloriozova, T. (1999) Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **39**, 666–670.
57. Xing, L., and Glen, R. C. (2002) Novel Methods for the Prediction of logP, pKa, and logD. *J. Chem. Inf. Comput. Sci.* **42**, 796–805.
58. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **44**, 170–178.
59. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **44**, 1708–1718.
60. Glen, R. C., Bender, A., Arnby, C. H., Carlsson, L., Boyer, S., and Smith, J. (2006) Circular Fingerprints: Flexible Molecular

Descriptors with Applications from Physical Chemistry to ADME. *IDrugs* **9**, 199–204.
61. Rodgers, S., Glen, R. C., and Bender, A. (2006) Characterizing Bitterness: Identification of Key Structural Features and Development of a Classification Model. *J. Chem. Inf. Model.* **46**, 569–576.
62. Cannon, E. O., Amini, A., Bender, A., Sternberg, M. J. E., Muggleton, S. H., Glen, R. C., and Mitchell, J. B. O. (2007) Support Vector Inductive Logic Programming Outperforms the Naive Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *J. Comput. Aided Mol. Des.* **21**, 269–280.
63. Faulon, J.-L., Visco, D. P., Jr., and Pophale, R. S. (2003) The Signature Molecular Descriptor. 1: Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **43**, 707–720.
64. Faulon, J.-L., Churchwell, C. J., and Visco, D. P., Jr. (2003) The Signature Molecular Descriptor. 2: Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **43**, 721–734.
65. Churchwell, C. J., Rintoul, M. D., Martin, S., Visco, D. P., Jr., Kotu, A., Larson, R. S., Sillerud, L. O., Brown, D. C., and Faulon, J. L. (2004) The Signature Molecular Descriptor. 3: Inverse-Quantitative Structure-Activity Relationship of ICAM-1 Inhibitory Peptides. *J. Mol. Graph. Model.* **22**, 263–273.
66. Avidon, V. V., and Leksina, L. A. (1974) Descriptor Language for the Analysis of Structural Similarity of Organic Compounds. *Nauchno. Tekhn. Inf., Ser. 2*, 22–25.
67. Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **25**, 64–73.
68. Horvath, D. (2001) High Throughput Conformational Sampling & Fuzzy Similarity Metrics: A Novel Approach to Similarity Searching and Focused Combinatorial Library Design and Its Role in the Drug Discovery Laboratory. in *Combinatorial Library Design and Evaluation: Principles, Software Tools and Applications* (Ghose, A., and Viswanadhan, V., Eds.), 429–472, Marcel Dekker, New York.
69. Horvath, D., and Jeandenans, C. (2003) Neighborhood Behavior of In Silico Structural Spaces with Respect to In Vitro Activity Spaces: A Novel Understanding of the Molecular Similarity Principle in the Context of Multiple Receptor Binding Profiles. *J. Chem. Inf. Comput. Sci.* **43**, 680–690.
70. Bonachera, F., Parent, B., Barbosa, F., Froloff, N., and Horvath, D. (2006) Fuzzy Tricentric Pharmacophore Fingerprints. 1: Topological Fuzzy Pharmacophore Triplets and Adapted Molecular Similarity Scoring Schemes. *J. Chem. Inf. Model.* **46**, 2457–2477.
71. Horvath, D., Bonachera, F., Solov'ev, V., Gaudin, C., and Varnek, A. (2007) Stochastic versus Stepwise Strategies for Quantitative Structure-Activity Relationship Generation – How Much Effort May the Mining for Successful QSAR Models Take? *J. Chem. Inf. Mod.* **47**, 927–939.
72. Schuffenhauer, A., Floersheim, P., Acklin, P., and Jacoby, E. (2003) Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **43**, 391–405.
73. MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada, *MOE, Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada*. www.chemcomp.com.
74. Franke, L., Byvatov, E., Werz, O., Steinhilber, D., Schneider, P., and Schneider, G. (2005) Extraction and Visualization of Potential Pharmacophore Points Using Support Vector Machines: Application to Ligand-Based Virtual Screening for COX-2 Inhibitors. *J. Med. Chem.* **48**, 6997–7004.
75. Byvatov, E., Sasse, B. C., Stark, H., and Schneider, G. (2005) From Virtual to Real Screening for D3 Dopamine Receptor Ligands. *Chembiochem.* **6**, 997–999.
76. Nilakantan, R., Bauman, N., Dixon, J. S., and Venkataraghavan, R. (1987) Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **27**, 82–85.
77. Kearsley, S. K., Sallamack, S., Fluder, E. M., Andose, J. D., Mosley, R. T., and Sheridan, R. P. (1996) Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **36**, 118–127.
78. Kuz'min, V. E., Muratov, E. N., Artemenko, A. G., Gorb, L. G., Qasim, M., and Leszczynski, J. (2008) The Effects of Characteristics of Substituents on Toxicity of the Nitroaromatics: HiT QSAR Study. *J. Comput. Aid. Mol. Des.* **22**, 747–759.
79. Kuz'min, V. E., Artemenko, A. G., Muratov, E. N., Lozitsky, V. P., Fedchuk, A. S., Lozitska, R. N., Boschenko, Y. A., and Gridina., T. L. (2005) The Hierarchical QSAR Technology for Effective Virtual Screening and Molecular Design of the

Promising Antiviral Compounds. *Antivir. Res.* **65**, A70–A71.
80. Bemis, G. W., and Murcko, M. A. (1996) The Properties of Known Drugs. 1: Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893.
81. Bemis, G. W., and Murcko, M. A. (1999) Properties of Known Drugs. 2: Side Chains. *J. Med. Chem.* **42**, 5095–5099.
82. Guener, O. F. (2000) *Pharmacophore Perception, Development, and Use in Drug Design*, Wiley-VCH Publishers, Weinheim.
83. Langer, T., and Hoffman, R. D. (2000) *Pharmacophores and Pharmacophore Searches*, Wiley-VCH Publishers, Weinheim.
84. Wang, J., Lai, L., and Tang, Y. (1999) Structural Features of Toxic Chemicals for Specific Toxicity. *J. Chem. Inf. Comput. Sci.* **39**, 1173–1189.
85. Kazius, J., McGuire, R., and Bursi, R. (2005) Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **48**, 312–320.
86. Cunningham, A. R., Rosenkranz, H. S., Zhang, Y. P., and Klopman, G. (1998) Identification of 'Genotoxic' and 'Non-Genotoxic' Alerts for Cancer in Mice: The Carcinogenic Potency Database. *Mutat. Res.* **398**, 1–17.
87. Klopman, G. (1984) Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **106**, 7315–7321.
88. Klopman, G., and Rosenkranz, H. S. (1984) Structural Requirements for the Mutagenicity of Environmental Nitroarenes. *Mutat. Res.* **126**, 227–238.
89. Klopman, G. (1985) Predicting Toxicity Through a Computer Automated Structure Evaluation Program. *Environ. Health Perspect.* **61**, 269–274.
90. Rosenkranz, H. S., Mitchell, C. S., and Klopman, G. (1985) Artificial Intelligence and Bayesian Decision Theory in the Prediction of Chemical Carcinogens. *Mutat. Res.* **150**, 1–11.
91. Klopman, G. (1992) MULTICASE. 1: A Hierarchical Computer Automated Structure Evaluation Program. *Quant. Struct. Act. Relat.* **11**, 176–184.
92. Klopman, G., and Rosenkranz, H. S. (1994) Approaches to SAR in Carcinogenesis and Mutagenesis: Prediction of Carcinogenicity/Mutagenicity Using MULTI-CASE. *Mutat. Res.* **305**, 33–46.
93. Klopman, G., Chakravarti, S. K., Harris, N., Ivanov, J., and Saiakhov, R. D. (2003) In-Silico Screening of High Production Volume Chemicals for Mutagenicity Using the MCASE QSAR Expert System. *SAR QSAR Environ. Res.* **14**, 165–180.
94. Sanderson, D. M., and Earnshaw, C. G. (1991) Computer Prediction of Possible Toxic Action from Chemical Structure: The DEREK System. *Hum. Exp. Toxicol.* **10**, 261–273.
95. Gombar, V. K., Enslein, K., Hart, J. B., Blake, B. W., and Borgstedt, H. H. (1991) Estimation of Maximum Tolerated Dose for Long-Term Bioassays from Acute Lethal Dose and Structure by QSAR. *Risk Anal.* **11**, 509–517.
96. Judson, P. N. (1992) QSAR and Expert Systems in the Prediction of Biological Activity. *Pestic. Sci.* **36**, 155–160.
97. Judson, P. N. (1994) Rule Induction for Systems Predicting Biological Activity. *J. Chem. Inf. Comput. Sci.* **34**, 148–153.
98. Barratt, M. D., and Rodford, R. A. (2001) The Computational Prediction of Toxicity. *Curr. Opin. Chem. Biol.* **5**, 383–388.
99. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **46**, 3–26.
100. Oprea, T. I. (2000) Property Distribution of Drug-Related Chemical Databases. *J. Comput. Aided Mol. Des.* **14**, 251–264.
101. Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002) Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **45**, 2615–2623.
102. Hann, M. M., and Oprea, T. I. (2004) Pursuing the Leadlikeness Concept in Pharmaceutical Research. *Curr. Opin. Chem. Biol.* **8**, 255–263.
103. Leo, A. J. (1993) Calculating log Poct from Structures. *Chem. Rev.* **93**, 1281–1306.
104. Tetko, I. V., and Livingstone, D. J. (2006) Rule-Based Systems to Predict Lipophilicity. in *Comprehensive Medicinal Chemistry II: In Silico Tools in ADMET* (Testa, B., and van de Waterbeemd, H., Eds.), 649–668, Elsevier, Oxford, UK.
105. Kubinyi, H. (1998) Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspect. Drug Discov. Des.* **9–11**, 225–252.
106. Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002) Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **45**, 4350–4358.
107. *Daylight Chemical Information Systems Inc.* http://www.daylight.com.
108. *Barnard Chemical Information Ltd.* http://www.bci.gb.com/.

109. *Tripos Inc.* http://www.tripos.com.
110. Jaccard, P. (1901) Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bull. Soc. Vaud. Sci. Nat.* **37**, 241–272.
111. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* **38**, 2894–2896.
112. Hull, R. D., Singh, S. B., Nachbar, R. B., Sheridan, R. P., Kearsley, S. K., and Fluder, E. M. (2001) Latent Semantic Structure Indexing (LaSSI) for Defining Chemical Similarity. *J. Med. Chem.* **44**, 1177–1184.
113. Hull, R. D., Fluder, E. M., Singh, S. B., Nachbar, R. B., Kearsley, S. K., and Sheridan, R. P. (2001) Chemical Similarity Searches Using Latent Semantic Structural Indexing (LaSSI) and Comparison to TOPOSIM. *J. Med. Chem.* **44**, 1185–1191.
114. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
115. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **44**, 1177–1185.
116. Ormerod, A., Willett, P., and Bawden, D. (1989) Comparison of Fragment Weighting Schemes for Substructural Analysis. *Quant. Struct. Act. Relat.* **8**, 115–129.
117. Godden, J. W., Furr, J. R., Xue, L., Stahura, F. L., and Bajorath, J. (2004) Molecular Similarity Analysis and Virtual Screening by Mapping of Consensus Positions in Binary-Transformed Chemical Descriptor Spaces with Variable Dimensionality. *J. Chem. Inf. Comput. Sci.* **44**, 21–29.
118. Godden, J. W., Stahura, F. L., and Bajorath, J. (2004) POT-DMC: A Virtual Screening Method for the Identification of Potent Hits. *J. Med. Chem.* **47**, 5608–5611.
119. Batista, J., Godden, J. W., and Bajorath, J. (2006) Assessment of Molecular Similarity from the Analysis of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **46**, 1937–1944.
120. Godden, J. W., and Bajorath, J. (2001) Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **41**, 1060–1066.
121. Batista, J., and Bajorath, J. (2007) Chemical Database Mining Through Entropy-Based Molecular Similarity Assessment of Randomly Generated Structural Fragment Populations. *J. Chem. Inf. Model.* **47**, 59–68.
122. Gute B. D., Basak S. C., Mills, D. and Hawkins, D. M. (2002) Tailored Similarity Spaces for the Prediction of Physicochemical Properties. *Internet Electron. J. Mol. Des.* **1**, 374–387.
123. Fourches, D. (2007) Modèles multiples en QSAR/QSPR: développement de nouvelles approches et leurs applications au design «in silico» de nouveaux extractants de métaux, aux propriétés ADMETox ainsi qu'à différentes activités biologiques de molécules organiques. Louis Pasteur University of Strasbourg, Strasbourg.
124. Guha, R., and VanDrie, J. H. (2008) Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **48**, 646–658.
125. Peltason, L., and Bajorath, J. (2007) SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **50**, 5571–5578.
126. Bonachera, F., and Horvath, D. (2008) Fuzzy Tricentric Pharmacophore Fingerprints. 2: Application of Topological Fuzzy Pharmacophore Triplets in Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **48**, 409–425.
127. Harper, G., Bradshaw, J., Gittins, J. C., Green, D. V. S., and Leach, A. R. (2001) The Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **41**, 1295–1300.
128. Geronikaki, A. A., Dearden, J. C., Filimonov, D., Galaeva, I., Garibova, T. L., Gloriozova, T., Krajneva, V., Lagunin, A., Macaev, F. Z., Molodavkin, G., Poroikov, V. V., Pogrebnoi, S. I., Shepeli, F., Voronina, T. A., Tsitlakidou, M., and Vlad, L. (2004) Design of New Cognition Enhancers: From Computer Prediction to Synthesis and Biological Evaluation. *J. Med. Chem.* **47**, 2870–2876.
129. Katritzky, A. R., Kuanar, M., Slavov, S., Dobchev, D. A., Fara, D. C., Karelson, M., Acree, W. E., Jr., Solov'ev, V. P., and Varnek, A. (2006) Correlation of Blood-Brain Penetration Using Structural Descriptors. *Bioorg. Med. Chem.* **14**, 4888–4917.
130. Katritzky, A. R., Dobchev, D. A., Fara, D. C., Hur, E., Tamm, K., Kurunczi, L., Karelson, M., Varnek, A., and Solov'ev, V. P. (2006) Skin Permeation Rate as a Function of Chemical Structure. *J. Med. Chem.* **49**, 3305–3314.
131. Katritzky, A. R., Kuanar, M., Fara, D. C., Karelson, M., Acree, W. E., Jr., Solov'ev, V. P., and Varnek, A. (2005) QSAR Modeling of

Blood: Air and Tissue: Air Partition Coefficients Using Theoretical Descriptors. *Bioorg. Med. Chem.* **13**, 6450–6463.
132. Mannhold, R., Rekker, R. F., Sonntag, C., ter Laak, A. M., Dross, K., and Polymeropoulos, E. E. (1995) Comparative Evaluation of the Predictive Power of Calculation Procedures for Molecular Lipophilicity. *J. Pharm. Sci.* **84**, 1410–1419.
133. Nys, G. G., and Rekker, R. F. (1973) Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules: Introduction of Hydrophobic Fragmental Constants (f-Values). *Eur. J. Med. Chem.* **8**, 521–535.
134. Leo, A., Jow, P. Y. C., Silipo, C., and Hansch, C. (1975) Calculation of Hydrophobic Constant (log P) from pi and f Constants. *J. Med. Chem.* **18**, 865–868.
135. Ghose, A. K., and Crippen, G. M. (1987) Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2: Modeling Dispersive and Hydrophobic Interactions. *J. Chem. Inf. Comput. Sci.* **27**, 21–35.
136. Ghose, A. K., and Crippen, G. M. (1986) Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I: Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **7**, 565–577.
137. Ghose, A. K., Pritchett, A., and Crippen, G. M. (1988) Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships III: Modeling Hydrophobic Interactions. *J. Comput. Chem.* **9**, 80–90.
138. Wildman, S. A., and Crippen, G. M. (1999) Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873.
139. Suzuki, T., and Kudo, Y. (1990) Automatic log P Estimation Based on Combined Additive Modeling Methods. *J. Comput. Aided. Mol. Des.* **4**, 155–198.
140. Convard, T., Dubost, J.-P., Le Solleu, H., and Kummer, E. (1994) SMILOGP: A Program for a Fast Evaluation of Theoretical log-p from the Smiles Code of a Molecule. *Quant. Struct. Act. Relat.* **13**, 34–37.
141. Wang, R., Gao, Y., and Lai, L. (2000) Calculating Partition Coefficient by Atom-Additive Method. *Perspect. Drug Discov. Des.* **19**, 47–66.
142. Wang, R., Fu, Y., and Lai, L. (1997) A New Atom-Additive Method for Calculating Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **37**, 615–621.
143. Balakin, K. V., Savchuk, N. P., and Tetko, I. V. (2006) In Silico Approaches to Prediction of Aqueous and DMSO Solubility of Drug-Like Compounds: Trends, Problems and Solutions. *Curr. Med. Chem.* **13**, 223–241.
144. Varnek, A., Kireeva, N., Tetko, I. V., Baskin, I. I., and Solov'ev, V. P. (2007) Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **47**, 1111–1122.
145. Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, O., Vayer, P., Solov'ev, V., Hoonakker, F., Tetko, I. V. and Marcou, G. (2008) ISIDA: Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided Drug Des.* **4**, 191–198.
146. Varnek, A., Fourches, D., Solov'ev, V., Klimchuk, O., Ouadi, A., and Billard, I. (2007) Successful "In Silico" Design of New Efficient Uranyl Binders. *Solvent Extr. Ion Exch.* **25**, 433–462.
147. Grubbs, F. E. (1969) Procedures for Detecting Outlying Observations in Samples. *Technometrics* **11**, 1–21.
148. Solov'ev, V. P., and Varnek, A. (2003) Anti-HIV Activity of HEPT, TIBO, and Cyclic Urea Derivatives: Structure-Property Studies, Focused Combinatorial Library Generation, and Hits Selection Using Substructural Molecular Fragments Method *J. Chem. Inf. Comput. Sci.* **43**, 1703–1719.
149. Fourches, D., Kireeva, N., Klimchuk, O., Marcou, G., Solov'ev, V., and Varnek, A. (2008) Computer-Aided Design of New Metal Binders. *Radiochim. Acta* **96**, 505–511.
150. Varnek, A., and Solov'ev, V. (2008) Quantitative Structure-Property Relationships in Solvent Extraction and Complexation of Metals. in *Ion Exchange and Solvent Extraction* (Sengupta, A. K., and Moyer, B.A., Eds.), Taylor and Francis, Philadelphia.
151. Horvath, D., Marcou, G., and Varnek A. (2009) Predicting the Predictability: A Unified Approach to the Applicability Domain Problem. *J. Chem. Inf. Model.* **49**, 1762–1776.
152. Hao Zhu, D. F., Varnek, A., Papa, E., Gramatica, P., Tetko, I.V., Öberg, T., Cherkasov, A., and Tropsha, A. (2008) Combinational QSAR Modeling of Chemical Toxicants Tested Against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* **48**, 766–784.
153. *MolConnZ*, version 4.05; eduSoft LC: Ashland, VA, 2003.
154. Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Öberg, T.,

Todeschini, R., Fourches, D., and Varnek, A. (2008) Critical Assessment of QSAR Models of Environmental Toxicity Against *Tetrahymena pyriformis*: Focusing On Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **48**, 1733–1746.

155. Huuskonen, J. (2000) Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777.
156. McElroy, N., and Jurs, P. (2001) Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **41**, 1237–1247.
157. Ran, Y., Jain, N., and Yalkowsky, S. (2001) Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **41**, 1208–1217.
158. Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A., and Giralt, F. (2001) A Fuzzy ARTMAP Based on Quantitative Structure-Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **41**, 1177–1207.
159. Downs, G., and Barnard, J. (2002) Clustering Methods and Their Uses in Computational Chemistry. *Rev. Comput. Chem.* **18**, 1–40.
160. Caruana, R. (1997) Multitask Learning. *Mach. Learn.* **28**, 41–75.
161. Varnek, A., Gaudin, C., Marcou, G.; Baskin, I., Pandey, A. K., and Tetko, I. V. (2009) Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *J. Chem. Inf. Model.* **49**, 133–144.
162. Tetko, I. V., Tanchuk, V. Y., and Villa, A. E. P. (2001) Prediction of n-Octanol/Water Partition Coefficients from PHYSPROP Database Using Artificial Neural Networks and E-State Indices. *J. Chem. Inf. Comput. Sci.* **41**, 1407–1421.
163. Efron, B. (1983) Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* **78**, 316–331.
164. Netzeva, T. I., Worth, A. P., Aldenberg, T., Benigni, R., Cronin, M. T. D., Gramatica, P., Jaworska, J. S., Kahn, S., Klopman, G., Marchant, C. A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G. Y., Perkins, R., Roberts, D. W., Schultz, T. W., Stanton, D. T., van de Sandt, J. J. M., Tong, W., Veith, G., and Yang, C. (2005) Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **33**, 155–173.
165. Jaworska, J., Nikolova-Jeliazkova, N., and Aldenberg, T. (2005) QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *Altern. Lab. Anim.* **33**, 445–459.
166. Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004) Similarity to Molecules in the Training Set is a Good Discriminator for Prediction Accuracy in QSAR. *Altern. Lab. Anim.* **44**, 1912–1928.
167. Fukumizu, K., and Watanabe, S. (1993) Probabililty Density Estimation by Regularization Method, in *Proceed. of the International Joint Conf. on Neural Networks*, pp 1727–1730.
168. Parzen, E. (1962) On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33**, 1065–1076.
169. Schioler, H., and Hartmann, U. (1992) Mapping Neural Network Derived from the Parzen Window Estimator. *Neural Netw.* **5**, 903–909.
170. Duda, R., and Hart, P. (1973) *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York.
171. van der Eijkel, G. C., Jan, van der Lubbe, J., and Backer, E. (1997) A Modulated Parzen-Windows Approach for Probability Density Estimation, in *Proceedings of the Second International Symposium on Advances in Intelligent Data Analysis, Reasoning about Data*, Springer-Verlag.
172. Kireeva, N. (2009) QSPR Ensemble Modeling of Stabilities of Metal-Ligand Complexes and Melting Point of Ionic Liquids. PhD thesis. Louis Pasteur University, Strasbourg.
173. Feuston, B. P., Chakravorty, S. J., Conway, J. F., Culberson, J. C., Forbes, J., Kraker, B., Lennon, P. A., Lindsley, C., McGaughey, G. B., Mosley, R., Sheridan, R. P., Valenciano, M., and Kearsley, S. K. (2005) Web Enabling Technology for the Design, Enumeration, Optimization and Tracking of Compound Libraries. *Curr. Top. Med. Chem.* **5**, 773–783.
174. Green, D. V., and Pickett, S. D. (2004) Methods for Library Design and Optimisation. *Mini Rev. Med. Chem.* **4**, 1067–1076.
175. Green, D. V. (2003) Virtual Screening of Virtual Libraries. *Prog. Med. Chem.* **41**, 61–97.
176. Varnek, A., Fourches, D., Solov'ev, V. P., Baulin, V. E., Turanov, A. N., Karandashev, V. K., Fara, D., and Katritzky, A. R. (2004) "In Silico" Design of New Uranyl Extractants Based on Phosphoryl-Containing

Podands: QSPR Studies, Generation and Screening of Virtual Combinatorial Library and Experimental Tests. *J. Chem. Inf. Comput. Sci.* **44**, 1365–1382.
177. Tetko, I. V. (2002) Neural Network Studies. 4: Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **42**, 717–728.
178. Vapnik, V. N. (1999) An Overview of Statistical Learning Theory. *IEEE Trans. Neural Netw.* **10**, 988–999.
179. Fujita, S. (1986) Description of Organic Reactions Based on Imaginary Transition Structures. 1: Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **26**, 205–212.
180. Jauffret, P., Tonnelier, C., Hanser, T., Kaufmann, G., and Wolff, R. (1990) Machine Learning of Generic Reactions: Toward an Advanced Comp uter Representation of Chemical Reactions. *Tetrahedron Comput. Methodol.* **3**, 335–349.
181. Vladutz, G. (1986) Modern Approaches to Chemical Reaction Searching, in *Approaches to Chemical Reaction Searching* (Willett, P., Ed.), 202–220, Gower, London.
182. Hoonakker, F. (2007) Graphes condensés de réeactions, applications à la recherche par similarité, la classification et la modélisation. Louis Pasteur University, Strasbourg.
183. Hoonakker, F., Lachiche, N., Varnek, A., and Wagner, A. (2009) Condensed Graph of Reaction: Considering a Chemical Reaction As one Single Pseudo Molecule. *The 19th International Conference on Inductive Logic Programming*. http://lsiit.u-strasbg.fr/Publications/2009/HLVW09.
184. Tetko, I. V., Bruneau, P., Mewes, H. -W., Rohrer, D. C., and Poda, G. I. (2006) Can We Estimate the Accuracy of ADMET Predictions? *Drug Discov. Today* **11**, 700–707.
185. Baskin, I. I., Halberstam, N. M., Artemenko, N. V., Palyulin, V. A., and Zefirov, N. S. (2003) NASAWIN – A Universal Software for QSPR/QSAR Studies. in *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions.* (Ford, M., Ed.), 260–263, Blackwell Publishing, Oxford, UK.
186. Halberstam, N. M. (2001) *Modeling Properties and Reactivity of Organic Compounds Using Artificial Neural Networks. Department of Chemistry*, Moscow State University, Moscow.
187. Leo, A. J., and Hoekman, D. (2000) Calculating log P (oct) with No Missing Fragments: The Problem of Estimating New Interaction Parameters. *Perspect. Drug. Discov. Des.* **18**, 19–38.
188. Honorio, K. M., Garratt, R. C., and Andricopulo, A. D. (2005) Hologram Quantitative Structure-Activity Relationships For A Series of Farnesoid X Receptor Activators. *Bioorg. Med. Chem. Lett.* **15**, 3119–3125.
189. Baskin, I. I., Skvortsova, M. I., Stankevich, I. V., and Zefirov, N. S. (1995) On the Basis of Invariants of Labeled Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **35**, 527–531.

Chapter 10

The Scaffold Tree: An Efficient Navigation in the Scaffold Universe

Peter Ertl, Ansgar Schuffenhauer, and Steffen Renner

Abstract

The Scaffold Tree algorithm (J Chem Inf Model 47:47–58, 2007) allows to organize large molecular data sets by arranging sets of molecules into a unique tree hierarchy based on their scaffolds, with scaffolds forming leaf nodes of such tree. The hierarchy is created by iterative removal of rings from more complex scaffolds using chemically meaningful set of rules, until a single, root ring is obtained. The classification is deterministic, data set independent, and scales linearly with the number of compounds included in the data set. In this review we summarize the basic principles of the Scaffold Tree methodology and review its applications, which appeared in recent medicinal chemistry literature, including the use of Scaffold Trees for visualization of large chemical data sets, compound clustering, and the identification of novel bioactive molecules. References to several computer programs, including also free tools available on the Internet, allowing to perform classification and visualization of molecules based on their scaffolds are also provided.

Key words: Scaffold, Ring, Scaffold classification, Scaffold hopping, Clustering, Chemical space, Scientific visualization

1. Introduction

The concept of a scaffold as the central core of a molecule is one of the basic concepts of medicinal chemistry. The scaffold gives a molecule its basic shape, determines whether the molecule is rigid or flexible, and keeps substituents in their positions. Global molecular properties, such as hydrophobicity or polarity, which are important for bioavailability and the fate of molecules in an organism, are also determined mainly by the composition of the scaffold. Electronic properties of the scaffold, including energies of frontier molecular orbitals, orbital distribution or atomic charges, determine reactivity of a molecule which in turn is responsible for its metabolic stability and toxicity. Scaffolds are often used to define classes of chemical compounds in patent claims. Scaffolds also play an important role in several modern

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_10, © Springer Science+Business Media, LLC 2011

techniques used in medicinal chemistry. One example is combinatorial chemistry and parallel synthesis, where various ring systems are used as the central cores of combinatorial libraries. Another popular technique applied in the drug discovery process is "scaffold hopping" [1, 2], where the goal is to "jump" in chemistry space, i.e., to discover a new bioactive structure starting from a known active compound via the modification of the central core.

Due to the importance of scaffolds and rings in medicinal chemistry, numerous publications exploring this topic from the cheminformatics point of view have appeared. In this short overview, we cannot cover all publications in this area; we selected therefore only several papers, which introduce novel methodology or present interesting results.

In an already classical paper of Bemis and Murcko [3], 5,120 known drugs were analyzed to identify the most common scaffolds. The authors found 2,506 different scaffolds, but without regard to atom type, hybridization, and bond order, half of the drugs in the set were described by only the 32 most frequently occurring scaffolds. Lipkus [4] presented a method for organizing ring systems based on their topology. Three simple descriptors that characterize separate aspects of ring topology were used in the study. This approach was applied to a database of 40,182 different rings that were derived from a comprehensive collection of rings extracted from the CAS Registry File. The study demonstrated that the distribution of rings is not compact but contains many significant voids. The same author analyzed also scaffolds extracted from more than 24 million molecules from the CAS Registry [5] in order to characterize scaffold diversity of organic chemistry. The distribution of scaffolds has been found "top heavy," with only a small number of scaffolds present frequently and with scaffold distribution conforming almost exactly to a power law. Wilkens et al. [6] described a recursive algorithm termed HierS for rapidly identifying all possible scaffolds within a set of molecules. Biological data were coupled to scaffolds by the inclusion of activity histograms, which indicate how the compounds in each scaffold class performed in previous high-throughput screening campaigns. Ertl et al. in their "Quest for the rings" paper [7] constructed all possible conjugated ring systems with one, two, and three 5- and 6-membered rings and identified bioactive regions of this large scaffold space by using self-organizing neural networks trained on a set of known bioactive molecules. The authors found that the bioactivity is sparsely distributed in the scaffold universe, forming only several relatively small "bioactivity islands." A similar study [8] enumerating 24,847 heteroaromatic systems with two rings predicted more than 3,000 ring systems of these as potentially synthetically feasible. Wester et al. [9] examined the distribution of scaffold topologies among several databases, including PubChem, WOMBAT, and Dictionary of Natural

Products. They found that more than 50% of examined structures exhibit only eight distinct topologies and that fused rings have slightly higher frequency in databases containing bioactive molecules.

All these papers agree that the number of scaffolds found either in existing collections of molecules or scaffolds that may be constructed in silico is huge. These conclusions stress the necessity to have a method that would allow easy analysis and visualization of distribution of scaffolds in large molecular collections.

2. The Scaffold Tree Method

The Scaffold Tree methodology organizes large molecular data sets by arranging molecules into a unique tree hierarchy based on their scaffolds. The scaffolds form leaf nodes of such hierarchy trees. The method is described in detail in the original paper [10], therefore here we provide only a short summary of the approach. The classification procedure begins by removing all terminal chains from molecules to obtain the central scaffold. Exocyclic double bonds and double bonds directly attached to the linkers are retained. From these scaffolds, rings are removed iteratively one by one until only one ring remains. Removal of a ring means that bonds and atoms which are part of the ring are removed excluding atoms and bonds which are part of any other ring. In addition, all exocyclic double bonds attached to the removed ring atoms are removed as well. If the removed ring was connected to the remaining scaffold by an acyclic linker, this linker is now a terminal side chain and is removed as well. If the removal of a ring would lead to a disconnected structure, this ring cannot be removed. This procedure is iteratively repeated, until a single "root" ring is obtained. Example of such iterative scaffold dissection is shown in Fig. 1. At each step, prioritization rules are applied to decide which ring to remove next. This leads to a unique, hierarchical classification of scaffolds, where each scaffold in the hierarchy is a well-defined chemical entity, which is contained in the original molecule as a substructure. Therefore, established procedures such as canonical SMILES [11] or InChI-Key [12] can be used to generate a canonical representation for each subscaffold in the hierarchy. The uniqueness, however, has its price: As each scaffold in the classification tree can have only one parent, one has to select the prioritization rules carefully in order to retain that part of the scaffold as parent, which characterizes it in a chemically intuitive way. The set of rules used to determine which rings should be removed next are fully described in [10]. The rules have been formulated in such a way that simple rings at

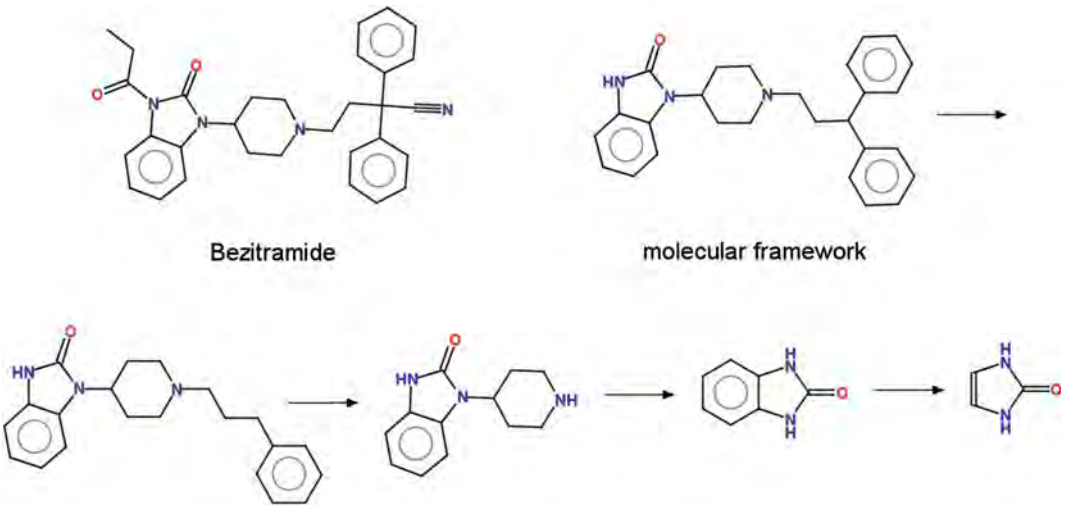


Fig. 1. Example of scaffold dissection of Bezitramide.

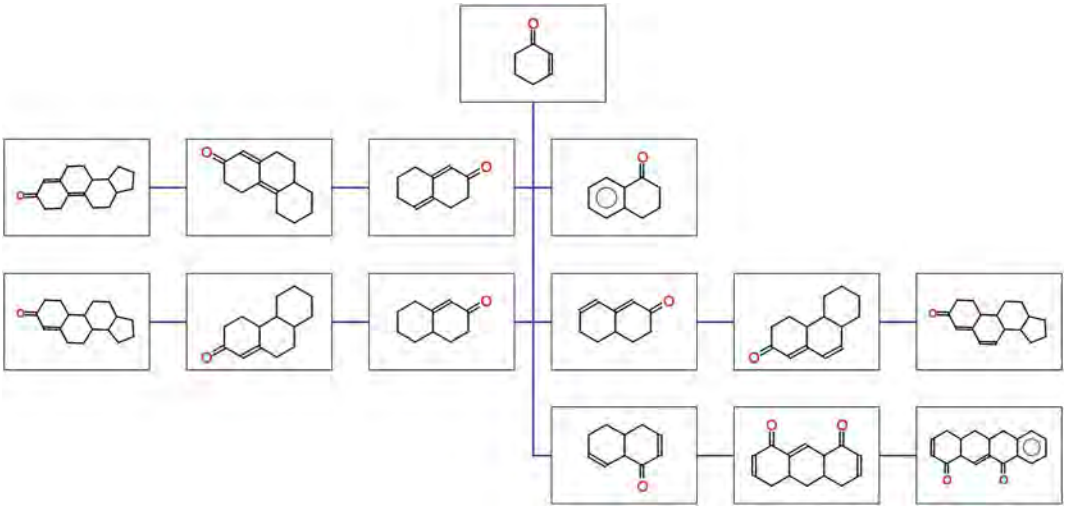


Fig. 2. One branch of a scaffold hierarchy with cyclohexenone as a root.

the periphery are removed first while the procedure tries to retain central and complex rings. In general, the goal of the method is to obtain a unique, chemically meaningful classification and not a classification with respect to pharmacophoric elements. Once the scaffold dissection is done for the whole set of molecules, the scaffold hierarchy of this set may be displayed (this is the final "Scaffold Tree" which gave the method its name). *See* Fig. 2 for an example of a single branch of such tree. The major advantage of this approach is that it offers a unique scaffold classification scheme that does not depend on a particular data set. Therefore,

scaffold hierarchies based on different data sets may be easily compared and merged.

Various rules may be used in the iterative scaffold dissection process to determine which ring should be removed next. In an early application of the Scaffold Tree method in the analysis of natural product structures [13], the frequency of occurrence of parent molecules in the database of natural products was one of the criteria used to decide which rings to retain, which provided some protection against selection of chemically nonintuitive subscaffolds. The disadvantage of this approach, however, is that the classification method is data set dependent, with the result that the further addition of structures to the data set would change the outcome of the classification. Therefore, in the further development of the Scaffold Tree approach we fine-tuned the prioritization rules in order to be able to dismiss the frequency as a prioritization criterion.

The use of biological activity as scaffold classification criteria can provide an advantage when studying particular data sets, for example results of high-throughput screening (HTS) or various collections of bioactive molecules. In such a scenario, the exhaustive identification of sets of scaffolds associated with the same biological activity might be advantageous over the more general data set independent scaffold hierarchy, defined by the chemical rules. Recently, a WOMBAT data set of molecules containing also biological activities was annotated using bioactivity as guiding criterion [14]. Branches connecting structurally complex and simple scaffolds having bioactivity on the same target were found at high frequency for the five major pharmaceutically relevant target classes. As already mentioned, however, this type of classification is not general and is valid only for a particular data set.

The most general analysis of scaffolds does not use any selection rules and simply takes into account all possible dissections. As already mentioned, such approach termed HierS has been described by Wilkens et al. [6] and used for compound clustering and analysis of HTS data. The complexity of the problem (the large number of possible scaffold hierarchies), however, makes visualization of all possible solutions very challenging.

3. Applications of the Scaffold Tree

3.1. Analysis and Visualization of Large Data Sets

The Scaffold Tree approach offers a unique, hierarchical classification of molecules based on their scaffolds, which is fast, deterministic, and data set independent. Each of the scaffolds in the hierarchy represents a viable chemical structure, which is a common substructure in all members of the scaffold class, instead of topological frameworks or reduced graphs (which are just abstract

mathematical objects) as in some other approaches. Since the method is data set independent, it is possible to classify compound sets individually and then overlay the resulted trees to detect in which chemical classes there is a overlap in the data sets.

The prioritization rules introduced in the method make it most likely that the scaffolds retained at higher hierarchy levels are chemically characteristic for the original molecules. There is no claim made, however, that the parts of the scaffolds are retained, which are responsible for the biological activity of the compound class. This would also generally not be possible since often the pharmacophoric features responsible for biologic activity are distributed over the whole molecule including the terminal side chains. However, if a specific ring system is used in a structure class because it presents the pharmacophoric side chains in the right geometric arrangement, then this ring system may be preserved as a common core in the whole class of active compounds, although it is not the pharmacophore itself. The method can also be used to "reverse-engineer" enumerated compound collections, such as offered by various companies selling compound libraries for screening, and identify the combinatorial library scaffolds, which have most likely been used.

The possibility to color branches of the final tree based on the concentration of bioactive molecules or the presence of molecules with specific type of activity makes this method particularly useful for presenting the analysis results to medicinal chemists. Since very large data sets can be processed with this method (we processed data sets with few millions of molecules), this approach is very useful for the analysis of results of HTS campaigns or the whole company archives. When, for example, Novartis acquired Chiron, the Scaffold Tree method was used to analyze and compare the compound archives of both companies. As a simple example of such visualization, results of Scaffold Tree analysis of the HTS screen for allosteric modulators of muscarinic M1 receptor from PubChem (bioassay 628 [15]) are shown in Fig. 3. In the hierarchy tree in this figure only such scaffolds are shown, which are present in at least 0.1% of the molecules in the complete data set and where at least 5% of the molecules assigned to this scaffold are active in the assay. Color intensity is used to indicate the fraction of active compounds containing this scaffold. This way to visualize scaffold hierarchy is very intuitive because color intensity immediately identifies those branches of the tree, which contain bioactive molecules. Typically, the scaffolds with a high fraction of actives would be those analyzed first for structure–activity relationship (SAR) and also checked regarding their intellectual property status. When analyzing data sets containing molecules with different types of activity, color in the Scaffold Tree may be used to identify regions of scaffold universe typical for different bioactivity types. In Fig. 4, an example of such visualization is shown. About 10,000

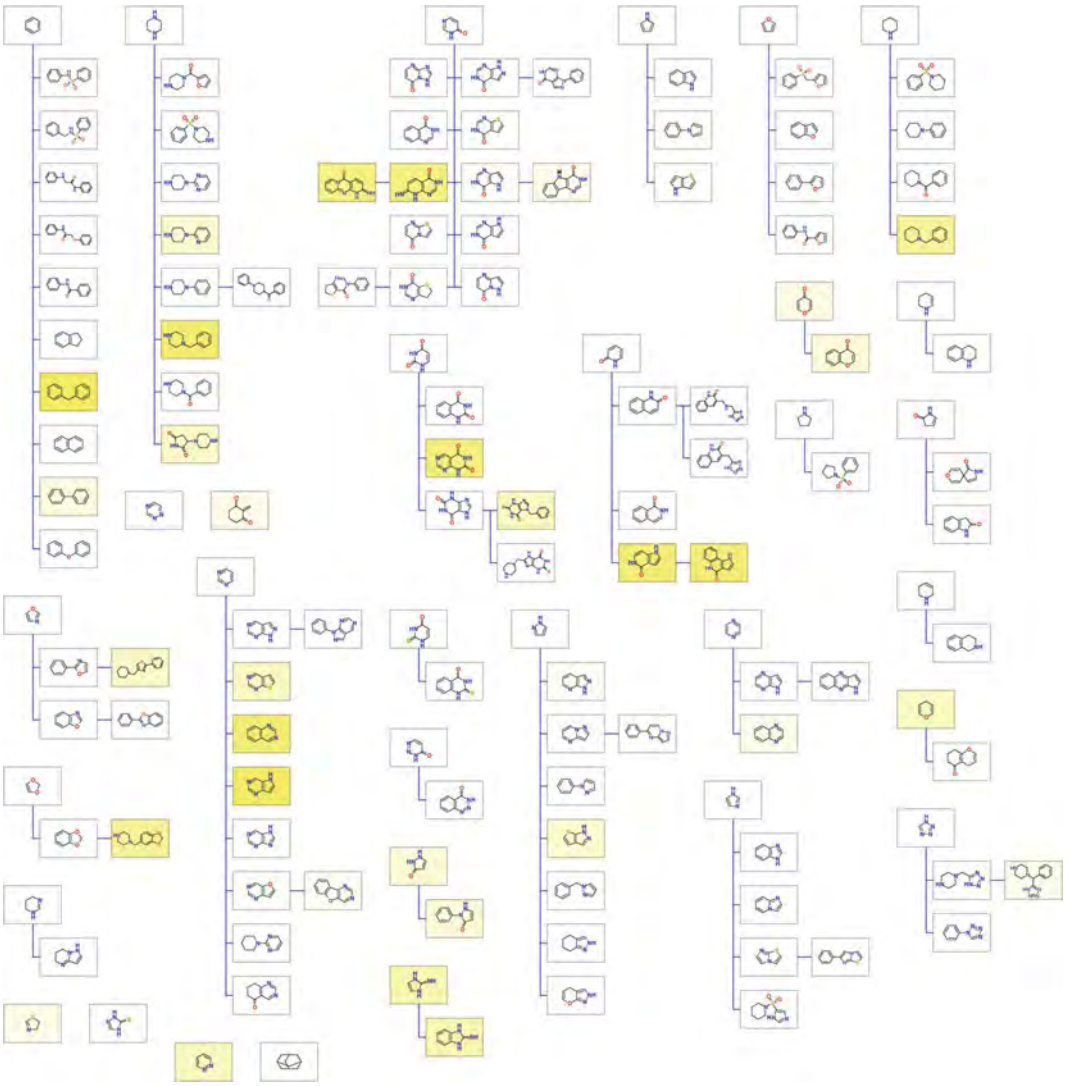


Fig. 3. Scaffold tree showing HTS results for allosteric modulators of M1 muscarinic receptor from PubChem. *Color intensity* represents portion of active molecules in this scaffold branch.

molecules from the GVK MedChem database [16] have been analyzed and the targets of molecules sharing a common scaffold are indicated under the respective scaffold by color coding.

The Scaffold Tree methodology was used also to perform a structural classification of natural products, SCONP [13]. The goal of the analysis was to provide an overview of the scaffolds that are present in natural products and could help to identify starting points for synthesis of novel bioactive structures. Structures from the Dictionary of Natural Products have been normalized and additionally, compounds that contained sugar moieties have been deglycosylated in silico prior to further analysis. Sugars

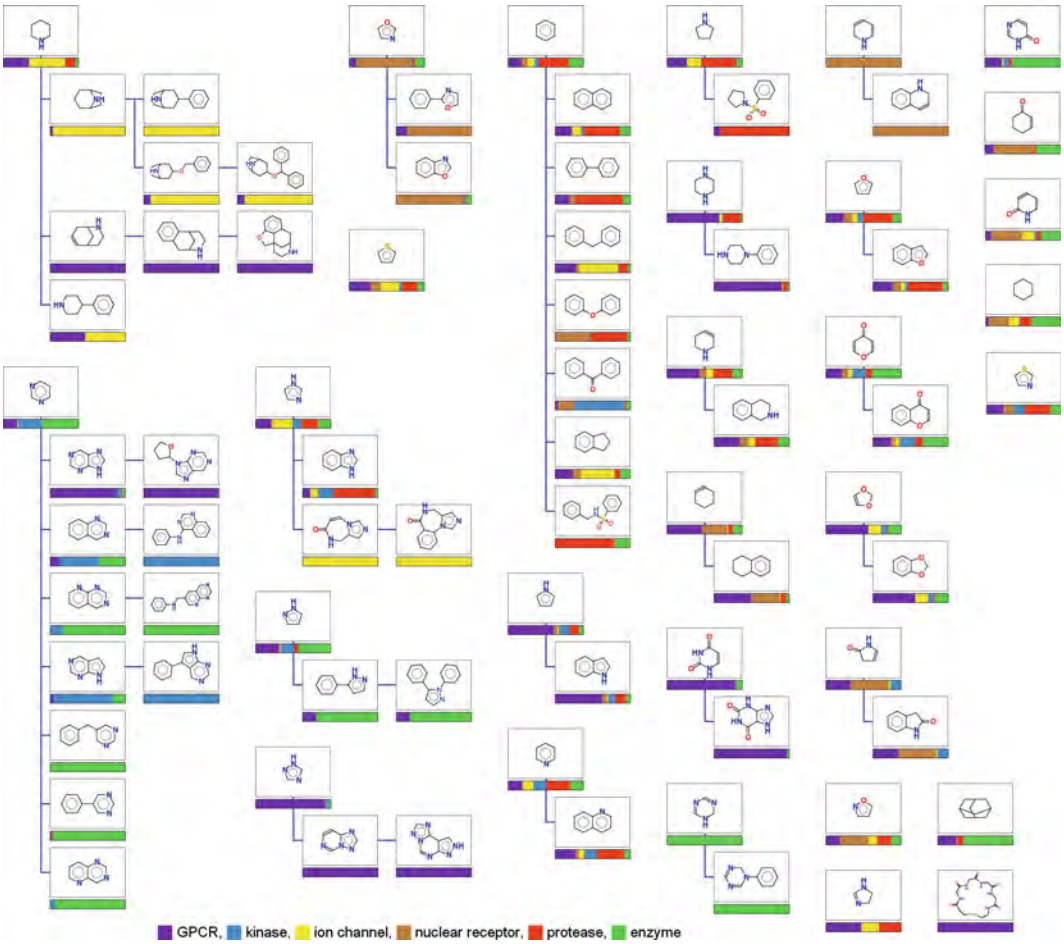


Fig. 4. Scaffold tree of representative molecules from the GVK database indicating also the preferred targets for molecules with this scaffold.

in natural products quite often serve only as solubilizing moieties and do not contribute to the biological action of the aglycons. As a result, all natural product structures could be represented in the form of a genealogical tree, the simplest and highest ranking elements of which were simple rings. Figure 5 shows a graphical representation of all natural product scaffolds that represent at least 0.2% of the records in the original database. This graph can give chemists a quick orientation on structurally related scaffolds. The validity of the approach was demonstrated by identification of a previously not described class of selective and potent inhibitors of 11ß-hydroxysteroid dehydrogenase type 1.

An analysis of the WOMBAT database using bioactivity-guided Scaffold Tree generation identified many scaffold branches with more than four active scaffolds for all five major protein classes of pharmaceutical relevance [14]. It can be assumed that

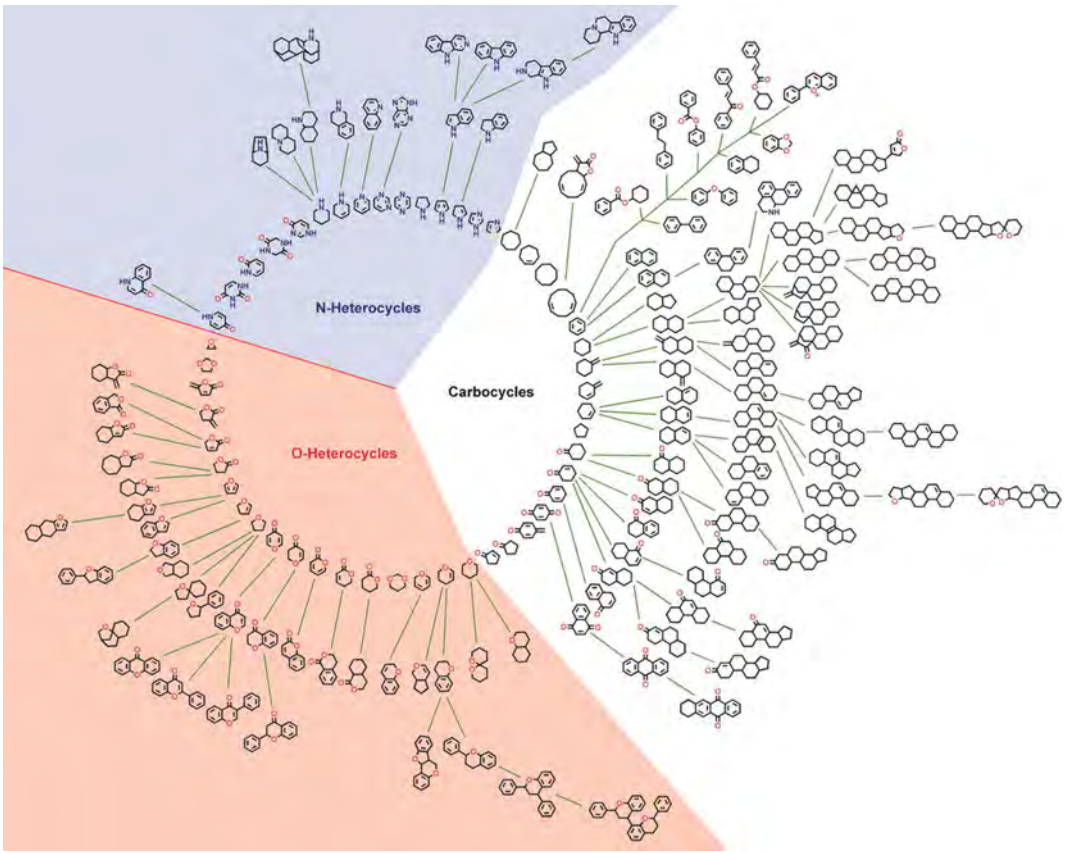


Fig. 5. Scaffold hierarchy of natural products. Figure from [13], Copyright (2005) National Academy of Sciences, USA.

this principle is widespread and will hold true for an even wider range of proteins.

3.2. Comparing Scaffold Tree-Based Partitioning with Other Clustering Methods

The Scaffold Tree method may be used with advantage also to obtain a hierarchical classification (or partitioning) of data sets based on chemical structures. At a given tree level, all structures sharing the same scaffold belong to the same partition. Scaffold Tree-based partitioning of molecular collections is entirely rule based, and the membership of a structure to a partition is not dependent on the overall composition of the data set. In contrast to this, most other clustering methods are based on unsupervised machine learning, which creates partitions depending on the particular data set. When used in the analysis of screening or bioactivity profiling data, the objectives for classification by structures are twofold: It is desirable to have as few partitions as possible whereby in each partition the molecules should be as similar as possible with respect to biological activity. These two objectives are competing. This can be easily seen by studying the extreme case solutions: one extreme case is to place all structures into one

partition, which has no homogeneity at all. The other extreme case is to place each structure into its individual partition and thereby maximizing the homogeneity of the partition at the price of creating only singletons. A Pareto analysis comparing several partitioning methods has been described in [17]. According to the Pareto ranking principle, a solution A of a multiobjective problem is considered superior to another solution B if A is superior to B in at least one objective and at least equivalent in all other objectives. In our case, the objectives to be considered are the number of classes generated, which should be as small as possible, and the homogeneity of the classes in the bioactivity space, which should be as high as possible. For data sets analyzed with a constant panel of assays, the homogeneity of a class could be determined as the average distance between all pairs of class members in the bioactivity space defined by the assay. No weighting of the objectives against each other (such as done in the Kelley penalty function [18]) has been applied. For each hierarchical partitioning method, cross-sections were made at several hierarchy levels computing thus for each partitioning method the trade-off curves between the two objectives. Three data sets have been analyzed, the NCI cancer cell-line screening, a proprietary data set with biochemical kinase assay profiling data, and a data with set safety pharmacology profiling data. On all three data sets, the rule based Scaffold Tree partitioning performed as good as those obtained with "classical" clustering methods such as the one available in Pipeline Pilot software, which is based on an algorithm similar to Optisim [19], or the divisive k-means clustering [20]. In contrast to these methods, the Scaffold Tree-based partitioning is deterministic whereas other clustering methods are not only depending on the whole data set, but also on the order of the records within the data set. This had been already shown by MacCuish [21] and confirmed by an analysis reported in [17]. By changing repeatedly, the order of molecules in the same data set and submitting it to the same clustering method and comparing the overlap of the solution using the adjusted Rand Index [22], it could be demonstrated that clustering can indeed be highly dependent on the order of records in the data set. In this aspect rule-based methods like the Scaffold Tree or the molecular equivalence numbers [23] have a clear advantage due to their deterministic nature, with the Scaffold Tree having additional benefit that its nodes are chemically viable entities.

3.3. Identification of New Bioactive Regions in the Scaffold Universe

Once the fragment hierarchy for a particular data set is generated, the resulting tree can provide a useful guideline for identifying novel bioactive regions in the scaffold universe. The technique called "brachiation" named according to the characteristic swinging locomotion of apes along the branches of "real" botanical trees may be applied here. This method is based on traversing

down the branches of the Scaffold Tree to identify simpler scaffolds sharing the important structural motifs with the original bioactive structure, or one jumps to the neighboring branches of known bioactive scaffolds with hope that these branches will also exhibit the same type of activity. Several examples of successful application of this technique have been published in the literature and are presented in this section and in Fig. 6.

The Scaffold Tree might be applied in various ways for the design of bioactive compounds and compound libraries. Biology oriented synthesis (BIOS) is one such approach based on the SCONP tree [13]. Since all scaffolds in the SCONP tree are derived from natural products, these scaffolds can be considered "prevalidated" by nature and represent privileged structural motifs for the design of general biology oriented libraries, not necessarily with a particular focus on a specific target. Several such libraries were designed in the Group of Prof. Waldmann, for an overview *see* [24]. For example, a library of a,ß-unsaturated d-lactones led to molecules modifying cell cycle progression **1** and the entry of vesicular stomatitis virus into cells **1**, **2** [25] and a library of spiroacetals gave modulators of tubulin cytoskeleton integrity **3** [26] (Fig. 6a).

Alternatively, one might also start from a natural product that already displayed the desired biological activity, e.g. hits resulting from HTS. Due to the structural complexity of natural products, it is often challenging to optimize such hits. Therefore, it is desirable to derive chemically more tractable compounds, retaining the biological activity. The Scaffold Tree provides a means to guide such a structural simplification. For example, the complex indole alkaloid Yohimbine **4** was found to have inhibitory activity on the phosphatase Cdc25A. Brachiation along the respective branch **8**–**12** of the SCONP tree of natural products identified the structurally simpler indoloquinolizidine **10** and indole scaffolds **11**. Both scaffolds represent a large group of natural products themselves, which suggests that biological relevance is retained during the brachiation process. Natural product-inspired compound collection based on these scaffolds gave potent hits on the phosphatase MptpB **5**–**7** [27] (Fig. 6b).

Scaffold Trees generated directly from HTS screening data or literature databases of compounds with annotated biological activities facilitate a more detailed analysis of biological activities within chemical space. In such trees, one can often find long branches or even clusters of branches of scaffolds having the same biological activity [14, 28]. Interestingly, within such clusters, one can often find scaffolds for which no molecules with activity of interest has been reported ("holes in bioactivity space"). This might be caused by an absence of the scaffold in the screened library, an unfavorable properties of sidechains, or the lack of annotation in literature databases. In all cases, such scaffolds are

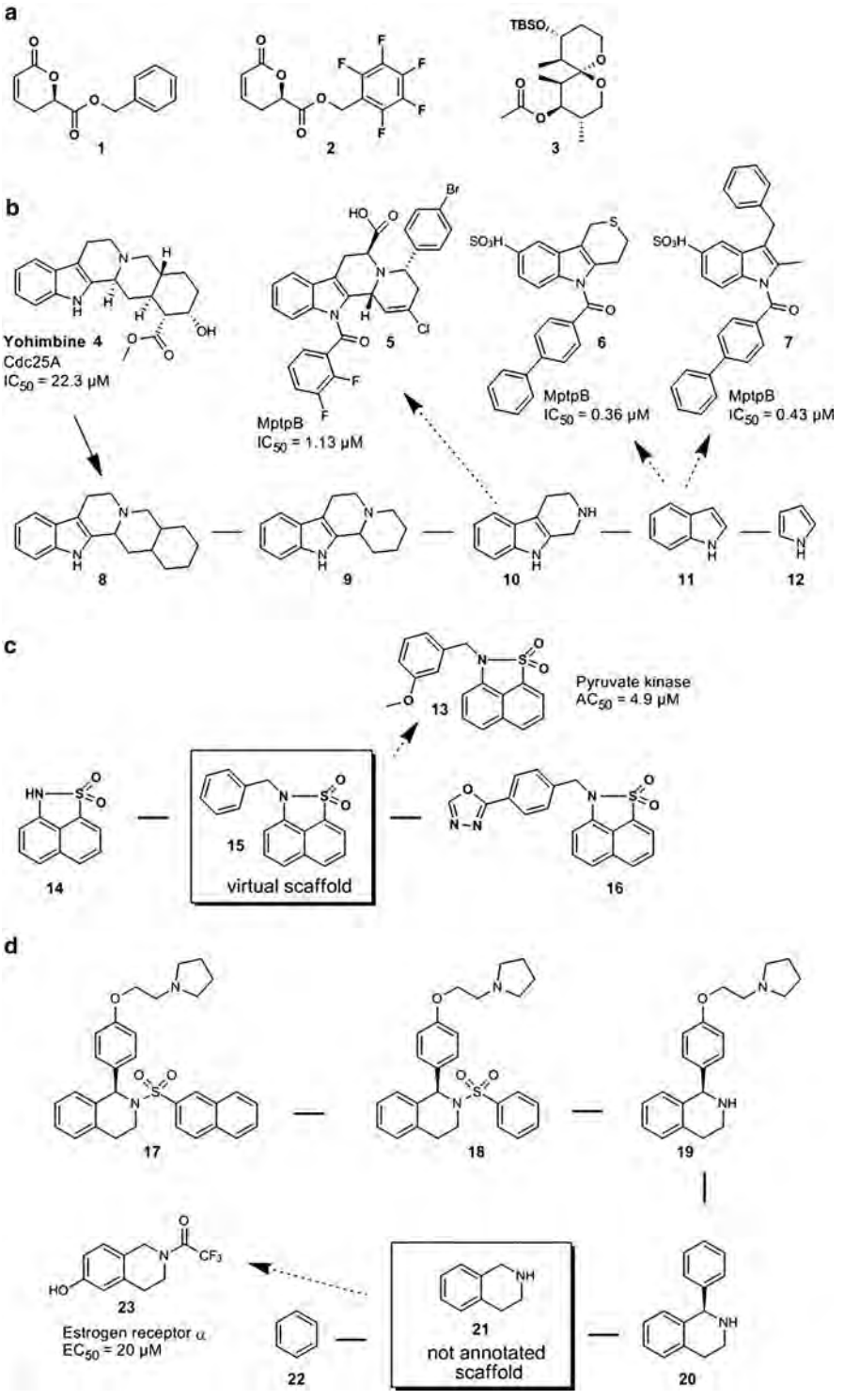


Fig. 6. Examples of bioactive molecules with novel scaffolds, designed using the Scaffold Tree methodology. *See* the main text for detailed description.

interesting candidates for follow-up experiments and represent promising starting points to identify novel scaffolds with a desired biological activity. Two examples are shown in Fig. 6c, d.

A Scaffold Hunter (*see* Subsection 4) analysis of the PubChem pyruvate kinase HTS data set revealed several interesting scaffolds, which were not present in the screening library, but were surrounded by active scaffolds [28]. Because the existence of such scaffolds is a result of the ring pruning procedure, they can be considered "virtual scaffolds," at least within the scope of the screening library. Here, **14** and **16** correspond to active members of the screening library from the same branch in the Scaffold Tree, connected by the "virtual scaffold" **15**. Molecules derived from **15** were found to be novel pyruvate kinase activators, e.g., **13**.

An analysis of the WOMBAT database using bioactivity-guided Scaffold Trees revealed many branches of scaffolds annotated with the same biological activity, distributed over the major target classes of pharmaceutical interest [14]. A branch with six scaffolds **17**–**22** representing compounds with activity on the estrogen receptor (ER-α) contained the tetrahydroisoquinoline **21** not annotated with ER-α activity. Synthesis of a small set of molecules based on this scaffold led to the active molecule **23** that was only about 100 times less active than the endogenous substrate estradiol (both measured in the same coactivator recruitment assay).

4. Software to Perform Scaffold Tree Analysis

The popularity of the Scaffold Tree approach prompted several commercial software vendors to implement this algorithm within their offerings. One example is the MOE package, where the Scaffold Tree is used as starting point for the detection of common scaffolds and subsequent R group analysis in sets of molecules [29]. Additionally, several programs that allow performing the scaffold analysis and clustering based either on the Scaffold Tree methodology or on similar approaches are available for free download or as free services on the Internet. The Molwind package [30] was developed by a team of scientists from Merck-Serono and made available for download as source code. The system dynamically generates molecule scaffold hierarchies from a set of inputted structures and displays them on a surface of a sphere, using the infrastructure of the NASA's "World Wind" sphere visualizer. For the organization of chemical data sets, a combination of Bemis–Murcko fragmentation, ring extraction, and iterative substructure searches is used, resulting in a hierarchy of fragments, which grow larger with increasing levels or depth. The program enables scientists to interactively browse chemical

compound spaces by changing between different levels of structural detail while maintaining relationships between similar compound classes. The already existing functionality and extendibility of the NASA's World Wind system provides an interesting basis for leveraging the interactivity offered by modern geospatial browsers in the area of research data exploration.

Another program for analysis and visualization of scaffolds available as source code is Scaffold Hunter [28]. Scaffold Hunter is a highly interactive computer-based tool for navigation in chemical space that enables intuitive recognition of complex structural relationships associated with bioactivity. The program reads compound structures and bioactivity data, generates scaffolds, correlates them in a hierarchical tree-like arrangement (Fig. 7), and annotates them with bioactivity. Brachiation along tree branches from structurally complex to simple scaffolds supports identification of bioactive structures.

The online clustering tool available at the Molinspiration web site [31] allows interactive clustering of data set using several similarity measures, including also clustering based on the Scaffold Tree hierarchy (Fig. 8). The online structure editor [32]

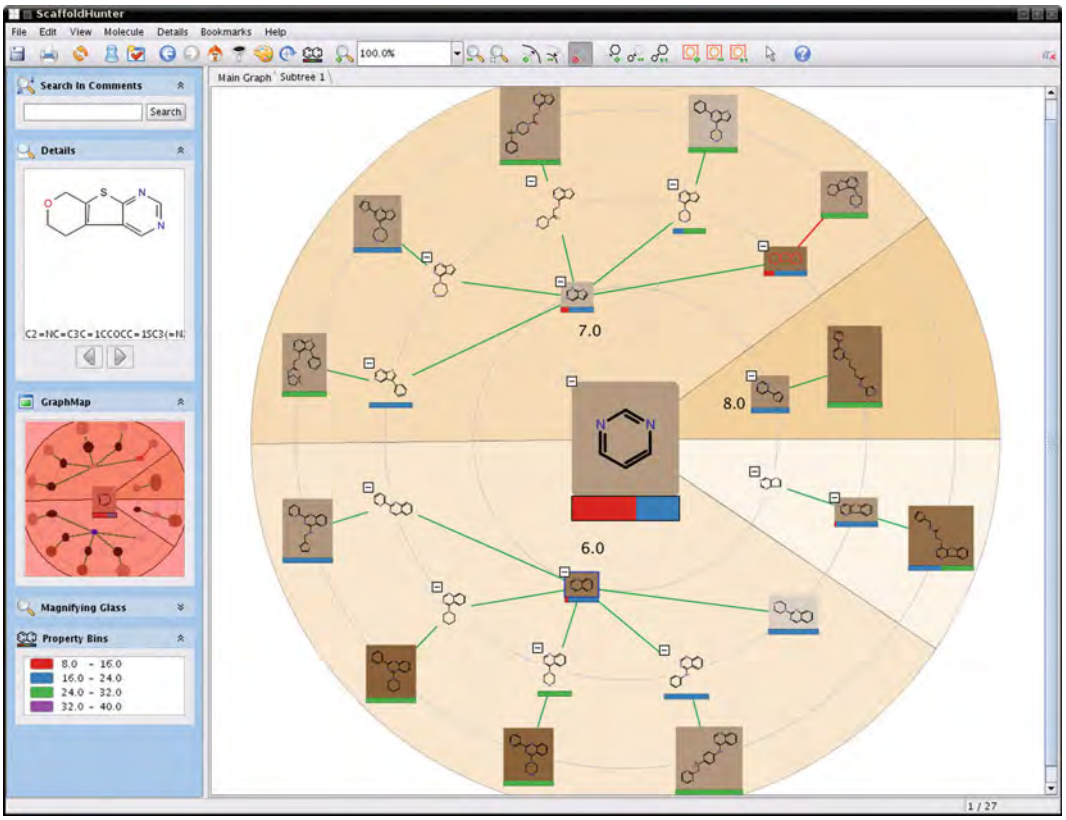


Fig. 7. Example of scaffold navigation using the Scaffold Hunter [28].

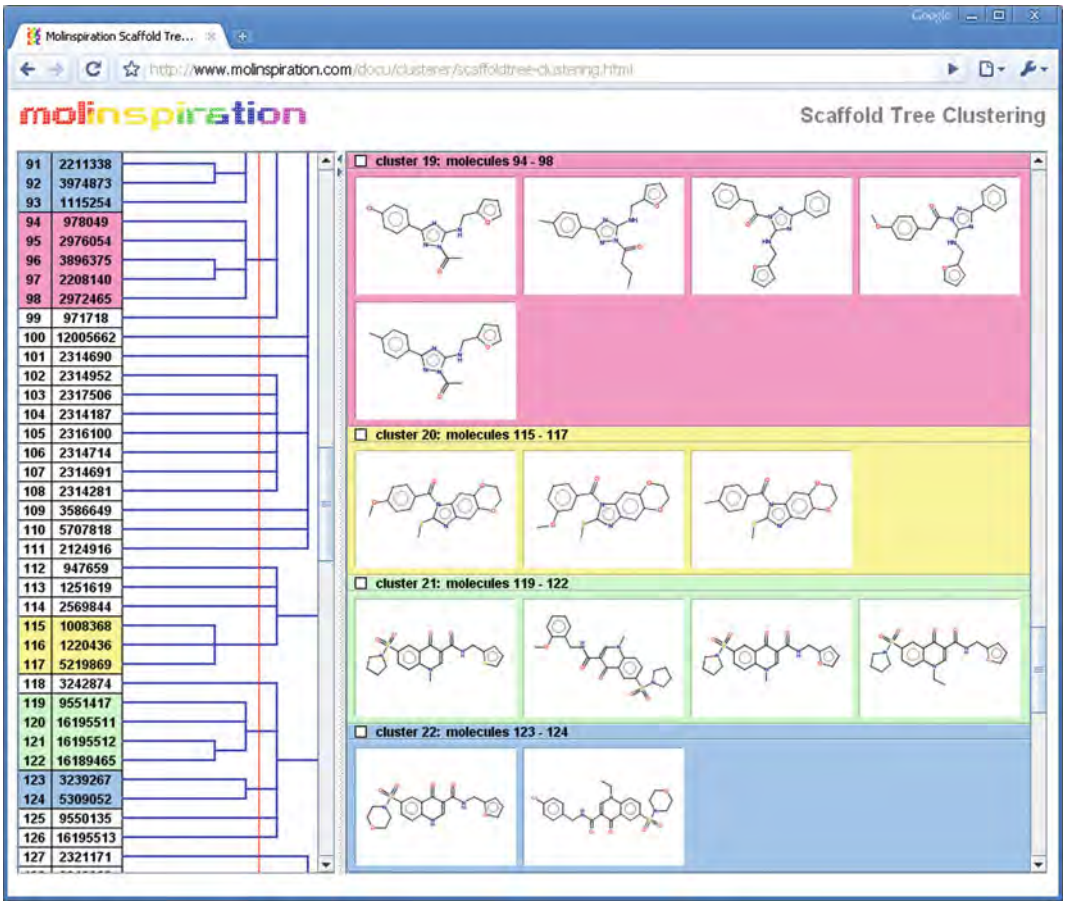


Fig. 8. Example of web-based Scaffold Tree clustering using the Molinspiration Clusterer [31].

powered by the CACTVS toolkit allows automatic dissection of molecules using the Scaffold Tree rules.

And one can hope that with the number of various cheminformatics and molecular processing tools available on the Internet increasing fast [33], the cheminformatics community will benefit also from other useful free tools for scaffold analysis and visualization, which appear in the future.

References

1. Brown, N. and Jacoby, E. (2006) On Scaffolds and Hopping in Medicinal Chemistry. *Mini Rev. Med. Chem.* **6**, 1217–1229.
2. Schneider, G., Schneider, P., and Renner, S. (2006) Scaffold-Hopping: How far can you Jump? *QSAR Comb. Sci.* **25**, 1162–1171.
3. Bemis, G. W. and Murcko, M. A. (1996) The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893.
4. Lipkus, A. (2001) Exploring Chemical Rings in a Simple Topological-Descriptor Space. *J. Chem. Inf. Comput. Sci.* **41**, 430–438.
5. Lipkus, A. H., Yuan, Q., Lucas, K. A., Funk, S. A., Bartelt, III, W. F., Schenck, R. J., and Trippe, A. J. (2008) Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **73**, 4443–4451.

6. Wilkens, S., Janes, J., and Su, A. (2005) HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs. *J. Med. Chem.* **48**, 3182–3193.
7. Ertl, P., Jelfs, S., Muehlbacher, J., Schuffenhauer, A., and Selzer, P. (2006) Quest for the Rings. In Silico Exploration of Ring Universe to Identify Novel Bioactive Heteroaromatic Scaffolds. *J. Med. Chem.* **49**, 4568–4573.
8. Pitt, W. R., Parry, D. M., Perry, B. G., and Groom, C. R. (2009) Heteroaromatic Rings of the Future. *J. Med. Chem.* **52**, 2952–2963.
9. Wester, M. J., Pollock, S. N., Coutsias, E. A., Allu, T. K., Muresan, S., and Oprea, T. I. (2008) Scaffold Topologies. 2. Analysis of Chemical Databases. *J. Chem. Inf. Model.* **48**, 1311–1324.
10. Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M. A., and Waldmann, H. (2007) The Scaffold Tree–Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **47**, 47–58.
11. Weininger, D. (1988) SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
12. http://www.iupac.org/inchi/.
13. Koch, M., Schuffenhauer, A., Scheck, M., Wetzel, S., Casaulta, M., Odermatt, A., Ertl, P., and Waldmann, H. (2005) Charting Biologically Relevant Chemical Space: A Structural Classification of Natural Products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **102**, 17272–17277.
14. Renner, S., van Otterlo, W. A. L., Dominguez Seoane, M., Moecklinghoff, S., Hofmann, B., Wetzel, S., Schuffenhauer, A., Ertl, P., Oprea, T. I., Steinhilber, D., Brunsveld, L., Rauh, D., and Waldmann H. (2009) Bioactivity-Guided Mapping and Navigation of Chemical Space. *Nature Chem. Biol.* **5**, 585–592.
15. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=628.
16. GVK MedChem Database, GVK Biosciences, http://www.gvkbio.com/.
17. Schuffenhauer, A., Brown, N., Ertl, P., Jenkins, J. L., Selzer, P., and Hamon, J. (2007) Clustering and Rule-Based Classifications of Chemical Structures Evaluated in the Biological Activity Space. *J. Chem. Inf. Model.* **47**, 325–336.
18. Kelley, L. A., Gardner, S. P., and Sutcliffe, M. J. (1996) An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally Related Subfamilies. *Protein Eng.* **9**, 1063–1065.
19. Clark, R. D. (1997) OptiSim: An Extended Dissimilarity Selection Method for Finding Diverse Representative Subsets. *J. Chem. Inf. Comput. Sci.* **37**, 1181–1188.
20. Engels, M. F. M., Gibbs, A. C., Jaeger, E. P., Verbinnen, D., Lobanov, V. S., and Agrafiotis, D. K. (2006) A Cluster-Based Strategy for Assessing the Overlap between Large Chemical Libraries and Its Application to a Recent Acquisition. *J. Chem. Inf. Model.* **46**, 2651–2660.
21. MacCuish, J., Nicolaou, C., and MacCuish, N. E. (2001) Ties in Proximity and Clustering Compounds. *J. Chem. Inf. Comput. Sci.* **41**, 134–146.
22. Hubert, L. and Arabie, P. (1985) Comparing Partitions. *J. Classif.* **2**, 193–218.
23. Xu, Y. J. and Johnson, M. (2002) Using Molecular Equivalence Numbers To Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **42**, 912–926.
24. Huebel, K., Lessmann, T., and Waldmann, H. (2008) Chemical Biology-Identification of Small Molecule Modulators of Cellular Activity by Natural Product Inspired Synthesis. *Chem. Soc. Rev.* **37**, 1361–1374.
25. Lessmann, T., Leuenberger, M., Menninger, S., Lopez-Canet, M., Müller, O., Hümmer, S., Bormann, J., Korn, K., Fava, E., Zerial M., Mayer, T. U., and Waldmann, H. (2007) Natural Product-Derived Modulators of Cell Cycle Progression and Viral Entry by Enantioselective Oxa Diels-Alder Reactions on the Solid Phase. *Chem. Biol.* **14**, 443–451.
26. Barun, O., Kumar, K., Sommer, S., Langerak, A., Mayer, T. U., and Waldmann, H. (2005) Natural Product-Guided Synthesis of a Spiroacetal Collection Reveals Modulators of Tubulin Cytoskeleton Integrity. *Eur. J. Org. Chem.* **22**, 4773–4788.
27. Nören-Müller, A., Reis-Correa, I., Prinz, H., Rosenbaum, C., Saxena, K., Schwalbe, H. J., Vestweber, D., Cagna, G., Schunk, S., Schwarz, O., Schiewe, H., and Waldmann, H. (2006) Discovery of Protein Phosphatase Inhibitor Classes by Biology-Oriented Synthesis. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 10606–10611.
28. Wetzel, S., Klein, K., Renner, S., Rauh, D., Oprea, T. I., Mutzel, P., and Waldmann, H. (2009) Interactive exploration of chemical space with Scaffold Hunter. *Nature Chem. Biol.* **5**, 581–583. http://sourceforge.net/projects/scaffoldhunter/.
29. Clark, A. M., and Labute, P. (2008) Detection and Assignment of Common Scaffolds in Project Databases of Lead Molecules. *J. Med. Chem.* **52**, 469–483.
30. http://molwind.sourceforge.net/.
31. http://www.molinspiration.com/docu/clusterer/.
32. http://www.xemistry.com/edit/frame.html.
33. Ertl, P. and Jelfs, S. (2007) Designing Drugs on the Internet? Free Web Tools and Services Supporting Medicinal Chemistry. *Curr. Top. Med. Chem.* **7**, 1491–1501.

Chapter 11

Pharmacophore-Based Virtual Screening

Dragos Horvath

Abstract

This chapter is a review of the most recent developments in the field of pharmacophore modeling, covering both methodology and application. Pharmacophore-based virtual screening is nowadays a mature technology, very well accepted in the medicinal chemistry laboratory. Nevertheless, like any empirical approach, it has specific limitations and efforts to improve the methodology are still ongoing. Fundamentally, the core idea of "stripping" functional groups of their actual chemical nature in order to classify them into very few pharmacophore types, according to their dominant physico-chemical features, is both the main advantage and the main drawback of pharmacophore modeling. The advantage is the one of simplicity – the complex nature of noncovalent ligand binding interactions is rendered intuitive and comprehensible by the human mind. Although computers are much better suited for comparisons of pharmacophore patterns, a chemist's intuition is primarily scaffold-oriented. Its underlying simplifications render pharmacophore modeling unable to provide perfect predictions of ligand binding propensities – not even if all its subsisting technical problems would be solved. Each step in pharmacophore modeling and exploitation has specific drawbacks: from insufficient or inaccurate conformational sampling to ambiguities in pharmacophore typing (mainly due to uncertainty regarding the tautomeric/protonation status of compounds), to computer time limitations in complex molecular overlay calculations, and to the choice of inappropriate anchoring points in active sites when ligand cocrystals structures are not available. Yet, imperfections notwithstanding, the approach is accurate enough in order to be practically useful and actually is the most used virtual screening technique in medicinal chemistry – notably for "scaffold hopping" approaches, allowing the discovery of new chemical classes carriers of a desired biological activity.

Key words: Pharmacophores, Ligand-based design, Structure-based design, Molecular overlay, Machine learning, Virtual screening, Conformational sampling

1. Introduction

Pharmacophores are defined [1] as "the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response." They

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_11, © Springer Science+Business Media, LLC 2011

represent a conceptual model aimed at describing structure-binding affinity relationships by means of a simple set of approximate rules-of-the-thumb. Chemistry cannot be, in practice, reduced to physics. Systems of very few atoms and, more important, very few degrees of freedom can be studied by means of rigorous quantum calculations, but, eventually, their results have to be "translated" back into chemical language – electrophilic/nucleophilic attacks, steric hindrance, etc. The human mind is notoriously unable to deal with particle wave functions.

Pharmacophores are, in a broad sense, the mental models and paradigms that form the basis of noncovalent chemistry. After understanding the three-dimensional nature of molecules and of the stereochemistry rules determining the preferred conformations, ligand binding to macromolecules was explained by the (oversimplified) key-and-lock paradigm [2] of shape complementarity. The nature of the noncovalent binding "forces" – electrostatic, hydrogen bonding, and dispersive contributions, including solvation/hydrophobic effects [3] – is however prohibitively complex. The reason for quoting the word "forces" above is that these are not fundamental (of which four are thought to govern our Universe – the nuclear/strong, the weak, boson-mediated, the electromagnetic and gravity). Steric "repulsion forces" are just a useful mental model to rationalize and "wrap up" the behavior of the electron clouds in interacting atom spheres. These are anything but "spheres" – yet, we tend (we have no choice but?) to think about them as such, and therefore we need to introduce a van der Waals corrective term, an extra hypothesis that is not required if the "sphere" model is dropped in favor of high level *ab initio* calculations. Much of modern chemistry happens at this "atom sphere" level of approximation, so pharmacophore modeling is certain to occupy a privileged position in the hearts and minds of medicinal chemists. The principle of functional group complementarity (cations interact favorably with anions, donors with acceptors, and hydrophobes among themselves) is an essential paradigm in modern medicinal chemistry.

Unsurprisingly, a query by the "pharmacophore" term in the Web of Knowledge [4] database returns several hundreds of citations per year. Simulation-based affinity predictions – flexible docking [5] or free energy perturbation simulations [6] – are typically too time-consuming to be of large-scale practical use (even though they are as well based on severe approximations of the physical reality, using empirical force field [7, 8] energy calculations).

Like in force field calculations, the first step in pharmacophore modeling is atom typing – classification of the atoms in terms of their nature and chemical environment, into predefined categories associated to a specific interaction behavior. Force field fitting is merely a much finer classification scheme, leading to force field

"types" associated to specific parameters describing the expected intensities of interaction. Pharmacophore typing typically does not go beyond a gross physico-chemical classification into "hydrophobes" (including or not the aromatic rings, which may be classified separately), "polar positives" (hydrogen bond donors and cations), and "polar negatives" (acceptors and anions). Unlike in force field typing, pharmacophore typing allows chemically different atoms being assigned to a same class (any lone-pair possessing heteroatom may in principle qualify as a hydrogen bond acceptor). Also, pharmacophore models do not provide any explicit characterization of the strength of interactions between features.

Next, a critical step in pharmacophore modeling is the conformational sampling of (a) known ligands (and known non-binders, essential negative examples in the machine learning process), to be used for pharmacophore extraction in so-called *ligand-based* approaches, and, (b) of all the candidate compounds from the electronic database that need to be confronted to the pharmacophore hypothesis, during the virtual screening. Conformational sampling is an extremely complex, multimodal optimization problem [9] that may require computer-intensive, massively parallel approaches for compounds exceeding a certain flexibility threshold (typically, several tens of rotatable bonds). Fortunately, drug-like compounds are less complex. Unfortunately, they are numerous. Reducing the conformational sampling time to a few seconds or, at most, minutes for each compound, using a simplified molecular force field for conformational strain energy estimations, does not guarantee the sampling of biologically relevant conformers. The sampling problem is far from being solved, as will be seen in §2.3.

The key step in pharmacophore modeling is obviously the construction of the pharmacophore hypothesis. A typical pharmacophore hypothesis (*see* Fig. 1) delimits a set of space regions (typically spheres) supposed to harbor functional groups of specified type when some low-energy conformer of the compound is optimally aligned with respect of it. The "optimal alignment" is the one allowing a maximum of spheres to be populated by corresponding groups – the relative importance of having each of them populated is an intrinsic estimate of the expected strength of the interaction it stands for.

These space regions are supposed to represent a map of interaction "hot spots" where favorable contacts to the protein site take place – which is certainly the case in *structure-based* approaches where these hot spots are picked from experimentally determined ligand–site cocrystals structure geometries. Structure-based ligand-free hypotheses (potential interaction points obtained by mapping of the empty active site) or ligand-based hypotheses (regrouping some consensus motif seen in active ligands, and therefore thought to be important for activity) are no longer sure

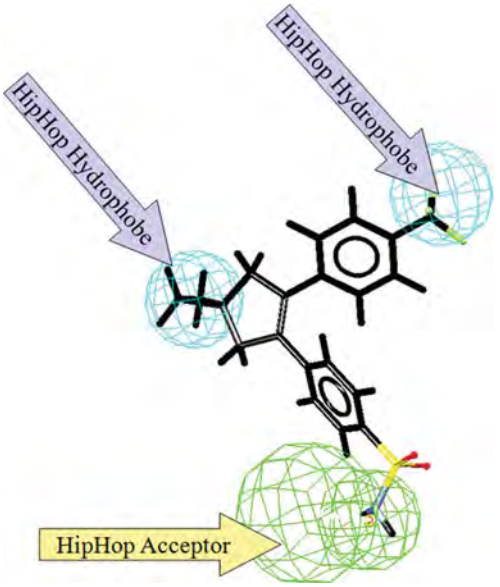


Fig. 1. Typical ligand-based pharmacophore model (extracted [31] by Catalyst [109]/HipHop on the basis of active Cox-2 inhibitors). Spheres delimit space zones supposed to harbor functional groups of indicated pharmacophore types, such as is the case of the pictured overlaid Cox-2 ligand. Note that hydrogen bonding interactions (and sometimes aromatic stacking too) are, in most commercial software packages, rendered directionally – they specify both a position for the ligand heavy atom and its polar hydrogen and/or lone pairs, or, respectively, a sphere for the expected protein partner atom.

to include the actual binding hot spots. Note that ligand-based hypotheses may be build either from an overlay model of (what are thought to be) calculated "bioactive" conformers of active compounds, or from machine-learning driven extraction of common patterns seen in the spatial distribution of pharmacophore groups in active ligands. In overlay-free models, common pattern extraction may as well (or perhaps better, when drastic geometry sampling artifacts hamper 3D modeling) be performed with topological distance values measuring separation between pharmacophore features. Methodologies exploiting such "topological pharmacophores" have been discussed elsewhere [10].

Eventually, the actual virtual screening is performed by confronting candidate ligands to the pharmacophore hypothesis. Some quantitative measure of match between a candidate conformer and the hypothesis needs to be defined beforehand (in case of ligand-based methods, this may – but need not – represent the objective function used to overlay the ligands, prior hypothesis generation). In overlay-free approaches, scoring is provided by summing-up the weighed contributions of key pharmacophore elements, like in a QSAR model. Ligands having at least one well-scoring conformation are then considered as the "virtual

hits" of the approach and should be subjected to synthesis and testing.

In terms of medicinal chemistry applications, the pharmacophore is often viewed as being complementary to the molecular scaffold. Scaffold hopping [11] became a central paradigm in drug design (*see* Fig. 2) – the quest of bioisosteric, topologically different structures, which nevertheless orient their interacting groups in space in a similar way to the starting compound and therefore display similar interactions with biological targets. Its importance stems from its ability to open new synthetic routes once that all the analogs around a given scaffold have been explored, to escape chemical space covered by scaffold-based patent applications, or to discover molecules with different pharmacokinetic properties but similar binding affinities with respect to the aimed target. Lead optimization is therefore alternatively oriented along two conceptually orthogonal research directions [12]: the sampling for various scaffolds compatible with a given pharmacophore pattern, and the sampling of various pharmacophore patterns that can be supported by a given scaffold.

Beyond scaffold hopping as defined above, new active molecules that features both a novel skeleton *and* a novel binding

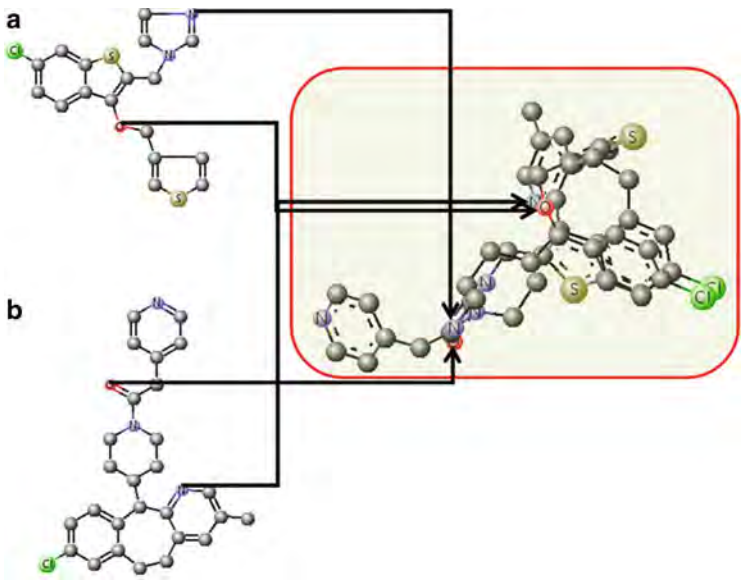


Fig. 2. A typical ligand-based scaffold-hopping scenario: both (**a**) and (**b**) are Farnesyl Transferase inhibitors although they are chemically very different compounds, not based on a common scaffold. The overlay model, however, evidences the pharmacophoric equivalence of certain functional groups. Besides the extensive overlap of hydrophobic/aromatic moieties, the hydrogen bond acceptors mapped by the arrows are equivalently distributed in space although they actually involve different heteroatoms (oxygens vs. pyridine/imidazole nitrogen atoms).

pattern to the target, hence corresponding to a completely original binding "paradigm," typically represent a major breakthrough in drug discovery. In addition to the above-noted advantages, these might, unlike the classical inhibitors matching the well-established pharmacophore, bind to different site pockets and therefore present different specificity profiles within the family of closely related targets. Obviously, simultaneous scaffold *and* pharmacophore hopping cannot be achieved with methods that learn a pharmacophore from a series of known binders, which do not provide any information on alternative binding modes. Therefore, this is not, strictly speaking, scaffold "hopping," for there is no reference scaffold to "hop" away from. The required information may only come from the protein structure, by designing novel putative binding pharmacophores, involving new anchoring points not exploited by known ligands. It is a potentially high-benefit, but certainly a high-risk approach, for not all the solvent-accessible hydrogen bond donors and acceptors in the active site are valid anchoring points (*see* discussion in Subsection 2.4.1).

Nowadays, pharmacophore-based virtual screening and modeling has reached maturity and has been extensively reviewed in past literature [13–25]. The goal of the present review is not to retrace the historical development of the domain, nor to provide any comprehensive list of commercial or academic software (programs are cited as mentioned in the reviewed articles). This contribution focuses on the very few latest years and the most interesting methodological developments and applications they brought. In the first section (Subsection 2), recently published technical issues will be presented, regrouped with respect to the strategies they address (structure vs. ligand-based, and, within this latter, overlay-based versus overlay-independent, respectively). Central key issues, common to all these approaches – such as conformational sampling – are discussed first.

Last but not least (Subsection 3), an overview of the latest pharmacophore virtual screening-based applications will be discussed, focusing only on articles presenting experimental validation of the found virtual hits.

2. An Overview of Latest Pharmacophore-Based Methodological Progress in Chemoinformatics

This chapter briefly reviews the latest methodological advances in pharmacophore modeling, covering first the general aspects (pharmacophore typing, conformational sampling), then structure-based and eventually ligand-based pharmacophore elucidation techniques.

2.1. Experimental Pharmacophore Detection

Although this is beyond the actual scope of this chemoinformatics-centric review, it is important to note that X-ray crystallography is not the only binding pharmacophore-elucidating technique. Nuclear Magnetic Resonance, monitoring either chemical shift changes of the resonance frequencies in the binding site, or relying on site/ligand Nuclear Overhauser Effects (NOE), may be used to extract the relative or even the absolute binding modes of ligands. In a recent contribution [26], it is shown that transfer-NOEs – transferring magnetization between the protons of two ligands competing for (weak reversible) binding to a common protein site, by means of the contact protons of the site – can be enhanced by reducing intraprotein spin diffusion (ideally, by deuterating the bulk protein, all in keeping the protons on the specific contact amino acids of the active site). A successful transfer of magnetization from the proton of a known "reference" ligand – of known binding mode – to a specific proton of another compound implies that the latter will occupy the same binding pocket. However, such methods are far from being routinely used in drug design, given the high degree of technological sophistication.

2.2. Pharmacophore Feature Detection (Typing): Does More Chemical Sense Make Better Pharmacophore Models?

Classically, pharmacophore typing is based on pragmatic rules: alkyl chains and halide groups are labeled as hydrophobes; aromatic rings may or may not be ordered into a specific "aromatic" category rather than joined to "ordinary" hydrophobes whereas polar groups are typed according to the number of attached polar hydrogens and/or lone pairs at expected protonation status. Also, the degree of resolution – flagging of individual atoms rather than considering "united" functional groups as pharmacophore feature carriers – is yet another empirical choice with respect to which various approaches largely differ. Some software tools allow the user to configure the pharmacophore typing schemes, others do not. To our knowledge, no two authors independently envisaging a pharmacophore typing scheme ever came up, by pure chance, with the same set of flagging rules. Some authors allow atoms to carry more than one pharmacophore flags (a carboxylate is both an "anion" and a "hydrogen bond acceptor") while others do not (in this case, in order to comply with the "one group – one type" policy, some amphiphilic donor-acceptor type needs to be introduced for, say, alcohol – OH groups).

However, different rule sets may nevertheless lead to a same pharmacophore typing result in straightforward situations in which there is, chemically speaking, not much place to argue that a carboxylate is an anion, the t-butyl group a hydrophobe and the ketone carbonyl an acceptor. The more complex the molecule, the likelier that any given rule-based flagging approach will fail for some functional groups. In particular, the uncertainty of the actual bioactive tautomeric form and the presence of multiple ionizable groups influencing each other's

pK_a values – such as the two aliphatic amine groups in piperazine rings, which are not simultaneously protonated at pH = 7, as rule-based flagging would suggest – are often sources of pharmacophore mistyping.

Recently, the introduction of a pharmacophore flagging scheme based on calculated estimations of pK_a values of ionizable groups [27] revealed that some of the observed activity cliffs [28, 29] – structurally very similar compound pairs with diverging activities – may be explained by subtle changes in pK_a values of ionizable groups, themselves translating into significant changes of relative population levels of the conjugated acidic/basic species. However, the application of this same strategy to generate fuzzy pharmacophore triplets as molecular descriptors used in QSAR studies [30] did not reveal any strategic advantage over classical rule-based flagging. QSAR build-up, in that context, can be assimilated to a pharmacophore elucidation approach, as it selects individual descriptors – pharmacophore triangles – seen to best correlate with observed activity. Or, pharmacophore learning – be it by QSAR-driven descriptor selection, inductive learning or overlay-based (*see* further chapters) – works best when the training set actives display as different a pharmacophore pattern as possible with respect to inactives. Therefore, the "best" pharmacophore flagging scheme, in this sense, is the one maximizing active versus inactive overall pattern dissimilarity, and *not* necessarily the physico-chemically most relevant one. The cited paper discusses one example in which the chemically wrong rule-based flagging scheme considered the benzodiazepine =N– in the 7-memebered ring as protonated. Yet, this wrong flag served as a primary marker of the – mainly active – set of compounds based on the benzodiazepine scaffold, by contrast to the mostly inactive ones, based on alternative scaffolds without fake cation.

The main problem [31] with pharmacophore typing scheme is, however, that the complex site–ligand interaction mechanisms cannot be rigorously understood in terms of some six or so functional group types. There is a universal consensus among all the flagging schemes – pK_a-based or not – on the issue that a carboxylate group (acceptor, anion) is pharmacophorically different, and thus not interchangeable with the hydrophobe $–CF_3$. Yet, it is also well known that the Cyclooxygenase II binding site easily accommodates $–CF_3$ groups in a carboxylate binding pocket. This is an example of the fundamental limitation of the pharmacophore concept, which, all its successes notwithstanding, represents an extremely sketchy and poor model of binding interactions. Also, the protonation state of a bound ligand may be, due to the influence of the binding pocket, different from the most populated state in solution. This effect cannot be taken into account by pure ligand-based approaches and is extremely difficult to model even if the structure of the binding pocket is known.

A recent [32] and potentially significant progress, in this respect, relies on the idea to completely give up pharmacophore typing according to a very restricted number of possible features but represent atoms by the more fine-grained, context-sensitive force field types [7] used in molecular mechanics parameterization of intramolecular interaction energies. Such typing would be too fine-grained if the "pharmacophore" hypothesis were to be formulated in terms of requiring atoms of strictly specified types at key positions of a ligand. However, the authors allow for fuzzy type matching – the substitution of a key atom of force field type t in an active by a different atom of type t' in a candidate molecule is hypothesized *not* to cause the loss of initial activity, in as far as types t and t' are related. The idea of allowing partial "cross-matching" of related types is not new (for example, in [27] aromatics and hydrophobes were assigned different flags but considered to be partially replaceable). The key element of originality here is the idea to objectively measure the degree of equivalence of t and t' based on how often they were actually seen to replace each other as anchoring points to a same "hot spot" of a protein site. Building a 3D "Flexophore" descriptor based on this flagging scheme, the authors find that "Flexophore descriptor detects active molecules despite chemical dissimilarity [from scaffold-hop benchmarking sets featuring no close analogs according to "classical" similarity scoring – N.A.] whereas the results for the screening of the complete data sets show enrichments comparable to chemical fingerprint descriptors." A further assessment [33] of Flexophore technology with respect to DUD molecules confirmed this scaffold-hopping ability.

The "ultimate" pharmacophore typing scheme is the one (apparently or effectively) abolishing pharmacophore types altogether and describing the chemical environment of functional groups by molecular field intensities, like in the now classical CoMFA [34] methodology. CoMFA and a plethora of related techniques do however continue to rely on force field typing schemes for partial charge and hydrophobe property assignments. Electrostatic fields may, however, be derived from quantum-mechanical calculations, and some authors recently[35] argued that given the steady increase of available computer power, this may no longer represent a computational bottleneck. This grants an effective independence from any empirical atom typing scheme and generates valid ligand overlays (*see* §2.6.1), but there are no compelling studies showing that such techniques are systematically superior to atom typing-based approaches. Indeed, a hyper-accurate, quantum level description of single or few conformers meant to serve as input for empirical overlay calculations does not make much physico-chemical sense: binding free energy is related to the Boltzmann ensemble properties of free and bound states, not to some empirical overlay score.

2.3. Progress in Conformational Sampling Techniques

Conformational sampling techniques are thus a core piece of the pharmacophore-based screening techniques, and still a hot research topic although a plethora of commercial and/or free software dealing with the problem is already available. Recent contributions to the topic [36–38] mainly deal with technical issues – intelligent conformer space coverage strategies accounting for fragment symmetry, multi-objective evolutionary strategies. Not fundamentally new, yet faster and more robust, the latest published approaches fail, however, to address the critical point of the accuracy of calculated strain energies, in order to enable specific selection of relevant geometries and hence minimize the odds of fortuitous pharmacophore matching. Certainly, the use of a state-of-the art force field [36] instead of the simple Catalyst energy function [38] may represent a good strategy. There are no definite rules concerning the maximal strain energy still tolerable in a bioactive conformer (and there never will be, for the physics of the ligand binding process is controlled by *free* energy, a costly parameter that cannot be routinely calculated as part of pharmacophore virtual screening). Some studies [39, 40] suggest something like ~1 kcal/mol of tolerable strain for each rotatable bond – way less than the typical 20–50 kcal/mol cutoffs used with commercial software, independently of ligand flexibility.

2.4. Structure-Based Pharmacophore Modeling

The terminology "structure-based" has been coined to refer to models generated from known ligand–site binding modes, or of empty protein binding sites featuring potential anchoring points, by contrast to "ligand-based" approaches where pharmacophore inference relies on information contained in the structures of known ligands (and, optionally, non-binders). Knowledge of the target macromolecule structure – and, potentially, of its key anchoring points used to bind known ligands – is the main advantage of structure-based approaches. Obviously, their use is limited to the cases where such information is available.

2.4.1. Extraction of Binding Pharmacophores from Empty Active Sites

This is the most difficult and risky structure-based design scenario since it relies on a relatively limited amount of experimental information: the plain protein structure (sometimes merely a homology model), some working hypothesis (or, at best, mutagenesis-based information) for localizing the active site and its key residues, and molecular simulation-generated information on the potentially flexible active site regions. Yet, this is also the potentially most interesting application, applicable to orphan targets, and was one of the first to be addressed by developers [41–43], and nowadays supported by most of the pharmacophore-building software suites. The active site of the protein is first probed by – typically – some "dry" hydrophobic and some charged probe spheres, to generate a map of "hot spots" where these probes witness the energetically most favorable interactions.

The "hot spots" are then clustered together and condensed into a most relevant set of pharmacophore feature spheres. Since there are countless possibilities to parameterize the monitored site energy maps and to condense the hot spots into a pharmacophore query, recent development is still ongoing in this field – the latest reported procedure [44] is based on the GRID [45] approach for hot spot mapping.

2.4.2. The Impact of Protein Flexibility in Structure-Based Pharmacophore Modeling

Typically, situations where structure-based design can be taken into consideration are rare and, if the prerequisite information concerning the active site is available at all, it may not go beyond a simple "snapshot" of the active site – empty or binding one ligand. Proteins are, however, flexible and may adapt to the incoming ligand. Therefore, the binding mode of a ligand may not be successfully inferred from the site geometry employed to bind another compound – or, for the matter, the empty site geometry. A recent, in-depth study [46] of structure-based pharmacophore extraction accounting for protein flexibility advocates the use of all the available active site geometries from various cocrystals (rather than using Molecular Dynamics-generated multiple protein geometries) in order to build a map of consensus interactions, present in >50% of the considered site conformations. These do not cover all the important anchoring points but have the merit of being entropically favored (they do not rely on the improbable event of having the active site adopt a very narrowly defined set of geometries). Molecules matching these key points have a good chance of being active, for they account for the "must-have" interactions with the rigid part of the active site. The remaining, flexible part of the active site is, by definition, able to "adapt" to the incoming ligands – generate new, unexpected favorable contacts, or, on the contrary, move away from the ligand and avoid bad contacts.

There are few macromolecular systems (such as the herein used Dihydrofolate Reductase) to boast the wealth of available ligand cocrystals structures needed for this ambitious study, which is likely to limit the interest of pharmaceutical industry for the methodology (a target known in such detail is no longer a "hot" issue for competitive drug development – but Molecular Dynamics-generated flexibility may be an alternative). Yet, results of the reported case study are quite encouraging. Pharmacophore hypotheses generated from multiple binding site conformers of human and *Pneumocystis carinii* DHFR were able to preferentially retrieve strong and weak DHFR binders over non-binders in virtual screening experiments. Furthermore, and surprisingly, they actually maintained their ability to select species-specific binding over promiscuous binders inhibiting the DHFR of multiple species although a loss of specificity is typically an expected consequence of flexible site modeling. This conclusion appears to

be strengthened by the fact that pharmacophore models derived from *Candida albicans* DHFR, with a significantly smaller set of different site conformers than used in the previous two cases, do actually lose species specificity in virtual screening.

In a further study [47], it was shown that this procedure appears to be even more successful when based on NMR-derived ensembles of geometries. Both models from the NMR ensemble and a collection of crystal structures were both able to discriminate known HIV-1p inhibitors from decoy molecules and displayed superior performance over models created from single conformations of the protein, but the NMR-based model appeared to be the most general yet accurate representation of the active site. This is in agreement with the observation that there is more structural variation between 28 structures in an NMR ensemble than 90 crystal structures bound to a variety of ligands. This work encourages the use of NMR models in structure-based design.

2.4.3. Detecting Meaningful Pharmacophore Anchoring Points in Protein Sites

Above-mentioned (Subsections 2.2 and 2.4) failures and limitations of the pharmacophore-based affinity scoring schemes should be a surprise to nobody. For the matter, docking procedures, which ultimately differ from pharmacophore matching tools only in terms of the hyperfine atom typing scheme in the force field/scoring function, are hardly better off in this respect. All the hypothesized "favorable" ligand–site interactions contribute, to the overall free energy, some small and highly context-sensitive increment, representing a small difference of several conflicting high-energy effects. Hydrogen bonding, for example, implies a favorable electrostatic interaction between a partially positively polarized ligand/site hydrogen atom and some electron lone pair of a partner heteroatom – at the cost of desolvating these partners, which previously formed (equally strong? stronger? weaker?) bridges to water molecules. The herein liberated waters may now connect among themselves in the bulk solvent – a favorable contribution to the ligand binding energy balance, which involves neither site nor ligand. Two H bonds (ligand–water and protein–water) are broken, two are formed (protein–ligand, water–water). Is the energy balance nil? Positive? Negative? It may be either way, depending on very subtle effects, such as the entropic aspects of all these interactions (would the protein–ligand bridge rigidify the implied protein side chain? Is the water molecule interacting with this side chain, in the uncomplexed state, restricted in its ability to form additional H bonds? ...). The hypothesis that protein–ligand H bonds are favorable may be statistically valid – they are more often favorable than not but this is of little help when the interaction of a peculiar ligand with a specified site has to be analyzed. Knowledge-based potentials are equally biased, for they are learned only on hand of complexes

with predominating favorable contacts – else, they would not have been available for X-ray snapshots.

In-depth molecular simulations [48] aimed at accurately capturing subtle entropic effects are in principle required to understand the actual contributions of individual contexts. Unfortunately, even if feasible, they are anything but routine tools to serve in high throughput drug discovery. No matter how sophisticated a simulation, it is likely that key interactions seen in all the crystal structures are not spontaneously scoring better energy/pharmacophore match increments than alternative, "fake" contacts – valid interactions in theory, but never seen to happen in practice.

Alternatively, machine learning from known protein–ligand complex structures can help to understand which specific hydrogen bonds, salt bridges, and hydrophobic contacts, out of all the possible ones that could be established by solvent-accessible atoms of a protein active site, are the strongest contributors to binding affinity. A recent study [49] showed that the prioritization of cavity atoms that should be targeted for ligand binding can be achieved by training machine learning algorithms with atom-based fingerprints of known ligand-binding pockets. The knowledge of hot spots for ligand binding is here used for focusing structure-based pharmacophore models.

The idea was taken one step further [50], from simple selection of significant interaction patterns, to a continuous weighing of considered interaction patterns (called "shims" in the original work) according to a partial-least-square analysis of the relationship between observed binding affinities and population status of each monitored interaction pattern. The herein calibrated "shim" contributions can be used as a target-specific correction of standard scoring functions of docking poses, or to learn affinity-predicting empirical models for homologue targets (for which crystal structure is not needed).

As always, machine learning is a powerful tool to evidence known patterns and to discover related ones, by means of limited extrapolation. While learning "hot spots" – the protein interaction sites actually used by some known ligands – from "cold" interaction sites (interaction opportunities not exploited by ligands), the method ignores whether the latter might be used by some not yet known ligands, or whether they are intrinsically inappropriate for binding. It was evidenced [51] that simple prioritization of the consensus features seen in multiple complexes of a protein with different ligands is enough to improve the enrichment scores of pharmacophore-based virtual screening experiments – sophisticated machine learning is not necessarily a must. Like always, consensus modeling consolidates what is known but minimizes the chance of discovery of radically new binding modes.

2.5. Progress in Automated Ligand Superposition and Ligand-Based Pharmacophore Elucidation

Since, in absence of an active site model, the putative ligand–site anchoring points remain unknown, ligand-based pharmacophore elucidation is typically based on some overlay model of active ligands, in order to evidence spatially conserved pharmacophore groups – and to assume (rightly or wrongly) that they are conserved *because* this is where the interaction with the active site occurs. Such alignment models are highly empirical and based on more than one shaky working hypothesis. In addition to the one cited above, lacking knowledge of the bioactive conformer forces ligand overlay users to assume that if two ligands possess a set of compatible geometries (that can be overlaid, pair-wise), then their bioactive conformers are somehow (!?) part of this set. Technically, the methods for searching (with respect to intra- and/or inter-molecular degrees of freedom, i.e., rigid/flexible overlays) and scoring the achieved overlay quality (typically some refined counting of the common features falling within a same pharmacophore feature sphere) vary widely, and – since based on empirical choices – none can be a priori considered superior. These approaches are rarely given a detailed description in literature, are often plagued by obscure empirical parameters set to some undocumented default values, and are often reinvented. Some of them are discussed in other reviews [14, 15, 17, 21, 23]. It is nevertheless important to point out that ligand-to-ligand fitting is intrinsically faster than ligand-to-site fitting (docking) – mainly because force field-based docking approaches include long-range interactions with protein atoms that are not necessarily in direct touch with the ligand. Using geometric and shape-matching techniques, authors [52] have actually managed to translate the docking problem into a ligand-to-ligand fit problem even in absence of an actual bound ligand, by rendering the site cavity as a virtual ligand. Latest contributions to the ligand overlay problem will be briefly discussed in the following.

2.5.1. Pharmacophore Field-Driven Superposition and 3D QSARs

The idea of using fuzzy pharmacophore "fields" [31, 53] of continuously decreasing intensity as a function of distance from their sources (typed atoms), instead of fixed-radius feature spheres containing or not the atoms of matching pharmacophore types, has been recently revisited several times [54–57]. The method assumes the optimal overlay to be the one maximizing the degree of overlap of corresponding pharmacophore fields (irrespectively whether they are entitled "fields [53]" or "Gaussian volumes [57]"), over the entire space surrounding the superimposed molecules. As such, it is less artifact-prone, for it does not demand any strong assumptions concerning the radii of classical pharmacophore feature spheres (instead, a fuzziness parameter may be used to smoothly control the tightness of match). Unsurprisingly, all the authors opted for the same empirical functional form of pharmacophore field intensity with respect to the distance to the

source – a Gaussian, very useful for analytical calculation of field overlap integrals. Typing and implementation details differ (notably, a pharmacophore hypothesis [54] is used as an overlay template, rather than the structure of a singled-out active compound [53]) – however, authors have seized the opportunity to use the pharmacophore field maps corresponding to optimal overlays as descriptors for 3D-QSAR training (which, in this case, can be assimilated to a pharmacophore elucidation process – but beware of surprises [31]!).

2.5.2. Pharmacophore Model-Driven Overlay

At first sight, this option makes little sense because ligand overlay is a prerequisite to pharmacophore model building, rather than being piloted by the latter. However, recent work [58] showed that it is possible to "co-evolve" overlay model and pharmacophore model (or receptor model, as termed in that work). Starting from some random alignment of random ligand conformers of known binders, a first – likely meaningless – pharmacophore model is established, then ligands are realigned with respect to the latter and stochastic iterations are pursued, featuring random changes in ligand geometry, position, considered receptor model points, etc. This approach constructs a receptor model setting "points" at space positions presumably occupied by the active site atoms, instead of feature spheres. Model scoring is not based on an overall ligand-to-ligand overlay score but on the goodness-of-match between the ligands and the emerging pharmacophore model, with a penalty term for pharmacophore model complexity. Thus, the approach seeks for the minimalistic pharmacophore model able to simultaneously accommodate all the ligands while simultaneously discovering how the ligands must be aligned. As a consequence, if the training set includes ligands covering several different binding modes, the minimalistic pharmacophore model should cover all of these modes, with each ligand matching its own specific subset of hot spots. The authors note the high enrichment ratios achieved by the superposition method even in comparison with procedures that exploit the protein crystal structure. However, ligand flexibility leading to a combinatorial explosion of the problem space volume is a critical issue in this approach – as in any other attempt to build ligand-based pharmacophore models.

2.5.3. Feature Pairing-based Overlay Algorithms

In pharmacophore field-based methods, six (3 rotational + 3 translational) degrees of freedom per overlaid ligand need to be exhaustively explored in order to pinpoint the optimal relative alignment. This may be costly but does not require any a priori matchmaking between corresponding functional groups, expected to be brought together during the overlay. Knowing beforehand which functional group of a ligand needs to be posed atop a given group of the reference compounds deterministically fixes the roto-translation required to minimize the Root-Mean-Square

(RMS) deviation between each reference group and its overlaid counterpart. Finding, in a molecule m, the functional group equivalent to a given feature in the reference compound M, is however not an easy exercise for a medicinal chemist (molecular overlay techniques were actually created to help visualizing the bioisosterism of apparently different groups) but can be successfully automated, as shown in a recent study [59]. Features in m and M need to be characterized by descriptors of the pharmacophore pattern surrounding them – then, features from m are putatively associated to the counterparts in M witnessing a similar surrounding pharmacophore pattern. This matching is sometimes far from being obvious. For example, if surrounding pharmacophore patterns are described only in terms of inter-feature distances, the algorithm ignores molecular chirality and likely suggests some impossible alignment. In the herein discussed implementation, this is not a fatal flaw, for alignment based on a pruned set of less stringent list of equivalent groups will be reattempted until some coherent pose is found. Near-optimal overlays can thus be effectively generated without needing to exhaustively explore a six-dimensional space.

However, this method is supposed to work best for compound pairs which do display a significant degree of pharmacophore feature similarity, thus containing enough unambiguously matching pharmacophore feature pairs. Pairs of marginally pharmacophorically similar compounds are better overlaid using the field approach. For example, consider a small molecule m and a much larger molecule M embedding a fragment m′, bioisosterically equivalent to m. The pharmacophore pattern descriptors of the features in m are, objectively, very different from those of the equivalent features in m′ (now surrounded by many more functional groups than there are in m). It is unlikely that they will outline m′ as equivalent group of m. Systematic rotations and translations, however, would eventually overlay m atop of m′ within M, and likely maximize local field covariance.

It is interesting to note that feature pairing-based alignment procedures can be, unlike field-based approaches, elegantly treated as a geometrical embedding problem [60] of a "hypermolecule." This is defined by the ensemble of ligands to be overlaid, where classical geometrical constraints are considered within each ligand (bond lengths, nonbonded exclusions). Atoms of different ligands do not see each other (are allowed to overlap – no inter-ligand non-bonded exclusions) and additional distance constraints are added to actually force equivalent pharmacophoric groups of different ligands to overlap. In fact, pharmacophore matching constraints can be refined in order to account both for spatial overlap of equivalent groups and the preservation of the directionality of hydrogen bonding or aromatic stacking interactions. Stochastic proximity embedding, an approach previously

used to conduct distance geometry-based conformational sampling, was now successfully generalized to generate conformers for each ligand, which overlap in terms of their equivalent groups. The elegance of the approach resides in the fact that intramolecular geometry and intermolecular alignment constrains are treated equivalently by the procedure. Unfortunately, the obtained geometries are merely guaranteed to be clash-free and feature more or less correct bond lengths and valence angles values – intramolecular energy is not explicitly being calculated in distance geometry approaches. Therefore, an energy-refinement postprocessing step is mandatory.

2.5.4. The Multiobjective Approach to Ligand Overlay

In flexible ligand overlay, the trade-off between goodness of fit (degree of overlap) and the acceptable strain energies of the overlaying conformers is a key empirical parameter, which is very difficult to set. How much strain energy is acceptable to increase, say, the field-based covariance score from 0.8 to 0.85? Alternatively, how much strain energy is acceptable to decrease the RMSD of the positions of equivalent functional groups by 0.5 Å? There is clearly no unambiguous way to weigh energy in kcal/mol against empirical overlay goodness scores (dimensionless correlation coefficients) or RMSD (Ångstrom). Strain energy and goodness of overlap are often conflicting objectives as energetically absurd geometry deformations may eventually allow any two compounds containing pharmacophorically equivalent groups to be perfectly overlaid. The total overlap volume is yet another independent overlay monitoring criterion – ensuring that the superimposed ligands are "squeezed" into a minimal volume. This may be mandatory for targets with very narrow active sites where suboptimal feature overlay and/or increased strain energies are the price to pay for fitting the site. Multiobjective optimization – the explicit dealing with multiple, potentially conflicting objectives to optimize, rather than forcefully selecting some empirically weighted linear combination thereof as the "ultimate" goodness criterion – is however a well-defined domain of numerical problem solving. In recent work [61], Pareto ranking methodologies were used to sample the space of possible partial overlays according to the three above-mentioned conflicting criteria of goodness-of-overlay, strain energy, and total overlay volume. The multiobjective framework leads to the identification of a family of plausible solutions, where each solution represents a different overlay involving different mappings between the molecules, and where the solutions taken together explore a range of different compromises in the objectives. The solutions are not ranked but are presented as equally valid compromises between three objectives, according to the principles of Pareto dominance. The approach also takes into account the chemical diversity of the solutions, thus ensuring that they represent a diverse range of structure–activity

hypotheses, which could be presented to a medicinal chemist for further consideration. It remains, however, unclear, how exactly to use this series of reasonable hypotheses for virtual screening. A consensus approach, picking only candidates matching all of these, may appear too restrictive since multiple hypotheses were generated to underline the fact that in diverse sets of binders, multiple anchoring patterns may coexist. The union of all features would, by contrast, generate a way too complex query retrieving only partial matches when confronted to real drug-like compounds in databases – unfortunately, the methodology is still under development and still has to prove the relevance of these partial matches.

2.5.5. Let Machine Learning Find Out How to Best Pilot Ligand Overlay!

Based on a training set of almost 70,000 "reference" overlays of protein–ligand complexes – generated on the basis of conserved amino acid residues in the protein sequences – authors [62] have recently shown that machine learning may predict what overlay template to use and which of the available software tools to employ, in order to maximize chances to reproduce, by means of ligand–ligand overlay, the "reference" ligand–ligand alignment. Random Forest models, trained using standard measures of ligand and protein similarity and Lipinski-related descriptors, are used for automatically selecting the reference ligand and overlay method maximizing the probability of reproducing the reference overlay deduced from X-ray structures (RMSD = 2 Å being the criteria for success). These model scores are highly predictive of overlay accuracy, and their use in template and method selection produces correct overlays in 57% of cases for 349 overlay ligands not used for training. The inclusion in the models of protein sequence similarity enables the use of templates bound to related protein structures, yielding useful results even for proteins having no available X-ray structures.

2.5.6. Alignment Rendered Simple: PhAST, the Linearized Pharmacophore Representation

Chemoinformaticians always envied bioinformaticians who deal with linear, diversity-restricted, and hence easy-to-align compounds, at the amino acid/nucleotide sequence level. Or, the pharmacophore paradigm allows boiling down the vast diversity of organic functional groups to a limited set of less than ten standard pharmacophore types, allowing the simplified representation of organic molecules as 2D graphs colored by pharmacophore types, and which can be thought of as the alternative to the standard bioinformatics "alphabets" of 20 amino acids or four nucleotides. The next logical step consists in linearizing this colored molecular graph to obtain a canonical sequence of pharmacophore types. Such sequences may then be aligned and compared according to bioinformatics-inspired metrics, dealing with simple operations of gap insertions instead of costly 3D rotations and field overlap scoring. Certainly, compression of the 2D graph

into a 1D sequence invariably triggers much loss of information. Nevertheless, the above-mentioned approach, recently [63] developed and tested, favorably compared to other virtual screening approaches in a retrospective study and identified two novel inhibitors of 5-lipoxygenase product formation.

2.6. Alignment-Free Ligand-Based Pharmacophore Elucidation

Ligand-based pharmacophore elucidation requires the detection of a consensus subset of features that are shared by all the actives using a common ensemble of anchoring points to the active site. Alignment, per se, is not a goal but merely a (rather costly and parameterization artifact-prone) way to outline this consensus subset of conserved features although it collaterally provides an intuitive depiction of the hypothesized binding mode. The classical form of alignment-free pharmacophore elucidation is QSAR-driven selection of 3D pharmacophore multiplets (pairs [64], triplets [24], quadruplets [65], etc.) that seem to correlate with activity, throughout a training set. The considered variables (binary yes/no population toggles, or fuzzy population levels of the multiplets) depend only on the considered conformers – see discussions in §2.3 – but not on their orientations. Unless fuzzy logics is used, geometry artifacts are however a key problem of this approach, which may be extremely sensitive with respect to the underlying conformational sampling procedure. Recently, authors [66] showed that "alignment-based alignment-free" approaches (using alignment-free multiplet descriptors, derived however from conformers that are able to smoothly align with geometries of other ligands) are better performers – doubtlessly due to the beneficial filtering stage represented by the alignment step, discarding many nonsense geometries.

2.6.1. QSAR-Driven Pharmacophore Elucidation?

Picking [67] relevant 3D pharmacophore triplets that correlate to the measured metabolic stability with respect to a given cytochrome (CYP 2D6) allows highlighting feature combinations that appear to facilitate binding to the enzyme in a way that aids substrate oxidation. However, it is highly unclear why pharmacophore signatures account for both the actual mechanistic aspect they were designed for – that is the ability of a substrate to be bound by the enzymatic site – and, in addition, predict whether the oxidation of the bound ligand will take place or not. True, no binding logically implies no metabolization – in this sense, filtering out the compounds that cannot possibly fit into the 2D6 active site is already significant. However, binding does not automatically imply metabolization – the article does not make a clear distinction between cytochrome inhibition and metabolic stability. QSAR, however, has long since been known to typically rely on correlations void of any causal background: the fact that metabolically unstable compounds tend to have a common signature in terms of 3D pharmacophore fingerprints does not yet imply that

the respective triplets are mechanistically involved in the studied property. In this particular case, selected pharmacophore features seem to match reasonably well known anchoring points in the CYP 2D6 site (which is still, per se, insufficient to explain stability). Other studies [30, 31] have clearly highlighted that valid pharmacophore descriptor-based QSARs are more likely to evidence typical signatures of actives versus inactives. Such signatures may converge toward the set of required binding interactions only if care is taken to use a highly diverse training set. In this context [30, 68, 69], it is worth mentioning that 3D information is not needed to highlight different pharmacophore pattern signatures – therefore, 2D pharmacophore fingerprints [10] are as useful in QSARs as their 3D counterparts (if not more useful, since void of conformational sampling artifacts).

2.6.2. Graph-Theoretical Approaches

Clique detection is a graph-theoretical approach mining for frequent common subgraphs within a set of graphs. A recent paper [70] reports an adaptation of this algorithm to deal with ligand-based pharmacophore elucidation while describing the pharmacophore pattern of each conformer in each ligand as a doubly annotated fully connected graph. Vertices are "colored" by the represented pharmacophore type while edges linking each feature to all the others are labeled by the corresponding Euclidean distances in a conformer. A first implementation mines all frequent cliques that are present in at least one of the conformers of each (or a portion of all) molecules. The second algorithm exploits the similarities among the different conformers and achieves significant speedups. These algorithms are able to scale to data sets with arbitrarily large number of conformers per molecule and identify multiple ligand binding modes or multiple binding sites of the target. A related approach is used by the MedSumo-Lig [71] software to calculate the match score of the graphs of pharmacophore triangles found in ligands.

2.6.3. Artificial Intelligence in Pharmacophore Elucidation

Alternatively, Inductive Logic Programming (ILP), a class of machine-learning methods able to describe relational data directly, can be used to express pharmacophores as a logical inference based on predicates related to the nature and relative distances of the key features entering the pharmacophore. In a recent publication [72], putative pharmacophore "hot spots" (features) are assigned to the energy minima of various Molecular Interaction Fields, generated by letting the molecule, in its current conformation, to interact with polar and respectively hydrophobic probes. Based on a list of actives – and, optionally, but highly desirable, inactive examples, the program seeks for inference rules like "A molecule is Active if (a) it possesses a hydrophobic interaction hot spot H1, AND (b) it possesses a positive charge interaction hot spot PC1, AND. . ., AND the distance between H1 and PC1 lays

between, AND. . .". Such rules may first be derived on the basis of a single active, then challenged to explain other actives: the higher the "coverage" – the generality – of the rule, the more trustworthy it can be considered. The method might actually deal with data sets of actives of different classes, which do not bind according to a common pharmacophore – and should successfully elucidate the characteristic pharmacophore of each class. Next, it has to be challenged with prediction of inactives in order to discard unspecific rules predicting all the molecules to be active. The great advantage of the method is the human-interpretable definition of the elucidated pharmacophore. Disadvantages, however, concern the likeness to extract artifactual, locally applicable rules which seem to hold for the training set due to the peculiar selection of included compounds, but are not genuine "rules of nature," like in classical QSAR [30]. However, the space of all the possible inference rules being huge – perhaps larger than the problem space of a typical attribute selection-based regression problem, it is difficult to assess the likeness of producing artifactual rules by inductive logic programming.

Even more recently [73], Inductive Logic Programming was used to generate pharmacophore-based sets of activity rules by means of known actives and inactives for several targets (this time using classical atom typing rules, not molecular interaction fields). However, these rules were not directly used as such to predict whether new structures are active or not but served to generate, for each molecule, a binary "rule compliance" fingerprint, in which bit number *i* is set for molecule M if M is predicted to be active according to rule *i*. Similarity screening was then reformulated as "Molecules with similar rule compliance fingerprints tend to have similar activities" and used to seek for neighbors of actives within a test dataset, containing hidden actives belonging to different chemical classes, in order to enforce "scaffold hopping." In parallel, classical similarity-based screening using state of the art scaffold hopping fingerprints CATS [74] was performed. Only the rule compliance fingerprints appeared to perform better than random selection. However, the herein reported comparison of a machine-learning based technique to an unsupervised similarity search is not very informative. Rule compliance fingerprints encode valuable chemical information learned from examples of actives and inactives. CATS fingerprints do not – they rely on the most basic assumption that activity similarity can be related to overall similar pharmacophore patterns (with no distinction made between actually binding groups and "bystanders," included to enhance physicochemical properties, or simply to make synthesis possible). Rule compliance fingerprint performance should rather be compared to a set of fingerprints based on multiple QSAR models, where locus *i* for molecule M contains the predicted activity of M according to, say, linear regression or Bayesian classifier model number *i*. Also,

the number of hidden actives per target was, in this study, extremely small – more in-depth benchmarking is required to evidence the real advantages of this otherwise elegant and promising inductive logic programming-based approach.

2.6.4. Partitioning-Based Pharmacophore Recognition

An interesting, recently reported approach [75] aims to define the common pharmacophore of a set of binders by detecting common k-point pharmacophore patterns (with an as large k as possible) in all the binders. In order to single out the potentially equivalent k-point pharmacophores, these are classified into cells based on the inter-feature distances, with k-point patterns belonging to a common cell (or within neighboring cells, to avoid binning artifacts) and found in all candidate ligands are considered for a final 3D overlay to assess the actual degree of spatial match (this is required, as the distance-driven binning scheme does not account for pattern chirality). A recursive distance partitioning algorithm is used to mine for reasonable pharmacophore classification schemes, leading to meaningful common pharmacophores.

2.7. Tuning of Pharmacophore-Based Virtual Screening Approaches: Efficiency Versus Performance

A recent study [76] on the impact of several screening parameters on the hit list quality of pharmacophore-based and shape-based virtual screening showed, intriguingly, that pushing the conformational sampling and the pharmacophore matching algorithm "too" far may prove detrimental to specificity – in the sense that careful searches for spurious poses of spurious conformations to match a given pharmacophore may eventually succeed. The authors recommend the use of CATALYST [77, 78] databases with a limit of maximally 50 generated conformers per ensemble and FAST generation algorithm combined with FAST database search as the default pharmacophore screening setup. Rising the number of considered conformers, or spending more time to match any given conformation to a pharmacophore increases the retrieval rate of actives though not specifically so: it becomes generally more likely to see some conformer of an inactive fitting, by chance, the pharmacophore models. Such observations are not new: cases reporting better performance of 3D pharmacophore fingerprints built on the basis of *fewer* conformers [79, 80] were already reported – then recently rediscovered [81] and deemed "surprising." Similarly – at least in appearance – parsimony is not only recommended in terms of conformer set sizes, but also in terms of monitored characteristics in pharmacophore fingerprints, where bit reduction/silencing may actually enhance performance [82]. However, the latter effect is not related to conformational sampling artifacts but due to intelligent focusing on the relevant fingerprint bits.

This is a fundamental problem, for it is tacitly assumed that considered conformer-to-pharmacophore model overlays represent the absolute optimum of a goodness-of-match score over all the possible degrees of freedom (intra- and inter-molecular) of the

problem. In reality, intramolecular degrees of freedom are "decoupled" from the pharmacophore match problem and treated separately – conformational sampling is performed "off line," not considering the pharmacophore to be matched and thus not favoring pharmacophore-matching geometries. Even so, too thorough an enumeration of possible geometries is likely to return conformers matching "everything" in terms of pharmacophore patterns. In as far their relative stability cannot be determined rigorously, in order to estimate relative populations in solution at room temperature, geometries that make physical sense are "hidden" amongst the many output by commercial software and cannot be foretold. Force field energy errors reach way beyond the characteristic $kT = 0.6$ kcal/mol – not to mention that, rigorously speaking, conformational *free* energies are the one controlling conformer population levels. For all these reasons, the unreliable intramolecular strain energies provided by the sampling tools are usually ignored when performing pharmacophore matches – all the geometries found within the geometry database are seen as equally valid hypotheses and the more of them, the likely that one will eventually match the pharmacophore query.

The good news highlighted by this study – in agreement with other assessments of the likelihood of commercial sampling software to generate, among others, the bioactive conformer [83–85] – is that, at least as far as the typical co-crystallized ligands are concerned, some reasonable geometry is often found among the best conformers (more of the best conformers needing to be considered for more flexible ligands). Therefore, actives adopting bioactive geometries close to default solution structures are more likely to be discovered when adopting a reduced conformer set strategy, for inactives then stand fewer chances to contribute some spuriously matching geometry. The bad news, however, is the inability to define some rigorous cutoff for the maximal excess strain energy that still allows a conformer to putatively qualify as bioactive geometry, for force-field-based energy values are way too imprecise. Therefore, the empirical cutoff in terms of optimal *numbers* of considered conformers makes no physical sense at all (and is likely problem- and software-dependent). Whether or not the bioactive conformer is part of the considered set and whether or not inactives will produce spuriously matching geometries (which should have been discarded if their actual energy levels would have been known) are two highly serendipitous aspects of pharmacophore-based virtual screening.

2.8. Pharmacophore Match: An Applicability Domain Definition for Docking?

Flexible docking [86–89] of putative ligands into an X-ray structure of the target site is expected to be the more powerful, comprehensive approach to modeling of ligand–site interactions because they allow – in principle – the discovery of any putative binding mechanism. Pharmacophore matching may only outline

whether a ligand matches a specified binding mode or not, whereas in docking the ligand has to "choose" the physically relevant binding mode out of all the possible, given an active site allowing for a certain number of contacts of specified nature. Reality is however different, as shown in a recent publication [90]. It either may be that sampling of possible poses and ligand conformers is insufficient – and the relevant one never gets enumerated, or that the correct one is being enumerated, but not ranked as the top stable one due to force field/scoring function artifacts. If all poses for each compound are passed through different pharmacophores generated from co-crystallized complexes, significantly larger enrichment factors (at a same selected subset size) are obtained based on the top-scoring passing pose of each compound. This, however, does not mean that pharmacophore models are, *per se*, a more realistic or more complete description of the binding site. They do perform better because they rely on additional experimental information – the ligand–site binding geometries seen in X-ray structures. Obviously, whenever a docking program generates a structure that is close to a known binding mode, it stands fair chances of having discovered an active. Poses which, however, do not match known binding modes, but nevertheless score well, are highly likely to represent scoring artifacts. Pharmacophore matching simply counts putative favorable contacts and would likely *not* having scored any better than docking if the list of considered contacts would have not only included the ones actually seen in experimental complexes, but all contacts that might have been possible within the active site. Why *the* specific hydrophobic contacts, hydrogen bonds or salt bridges seen in experimental structures seem to contribute much more to complex stability than other, theoretically at least as valid hydrophobic contacts, hydrogen bonds or salt bridges show the intrinsic limitation of both and conceptually quite related pharmacophore and scoring function/force field typing methodologies. All pharmacophore contacts appear equal in the virtual screening methodology, but some (the "native") are "more equal than the others." The underlying reason is likely hidden within highly subtle flexibility-and solvent-induced enthalpic and entropic effects that cannot be modeled at the typical resolution scale of the pharmacophore paradigm. The entire, booming, research field of docking interaction fingerprints [91–94] came to life in order to allow a quick discrimination of native-like poses from exotic, likely artifactual ones. Pharmacophore matching or fingerprint analysis, both these approaches aim to restrict the Applicability Domain [95–97] of docking – which can be viewed as a complex, nonlinear Quantitative Structure–Activity Relationship [98–100] – to the neighborhood of the experimentally known realm. However, the price for safely high hit rates is the risk of discovering only

"dull" analogs of known binding modes while filtering out the rare, but original ligands that actually do bind differently.

3. Recent Applications of Pharmacophore-Based Virtual Screening

Medicinal chemistry publications of the latest years abound in applications of pharmacophore-based virtual screening approaches used to discover new bioactive compounds. As noted in Introduction, this is now a mature chemoinformatics approach, intuitive and representing an apparently satisfactory trade-off between physicochemical rigor and computational effort. Categorizing interactions in "hydrophobic," "hydrogen bonds" and "salt bridges" is apparently good enough to let us explain, in many cases, the mechanism of ligand–site binding.

Unfortunately, we have no reliable records of failed pharmacophore-based virtual screening attempts, which never made it to the press. Obviously, the reasons for these failures are unknown – they may range from the simple absence of active compounds matching the pharmacophore pattern (not a methodological problem) to "drowning" of actual actives amongst too many false positives. At a technical level, they may be due to improper atom typing, unsatisfactory conformational sampling, biased machine learning/pharmacophore elucidation or sheer inappropriateness of the pharmacophore paradigm in that context (very flexible protein targets, atypical interactions requiring quantum-mechanical modeling, etc.).

Another less than welcome peculiarity of the recent literature concerning pharmacophore models is the wealth of publications, which apply existing virtual screening techniques – in most cases, a proper use thereof being reported, and very often involving a cascade of quick empirical filters followed by more rigorous docking approaches – but stop short of experimental validation of the selected hits [101–106]. The reported work may well be commendable and valid – yet, what was it done for if there is no interest from experimentalist groups to assess the activity of virtual hits and to continue developing them into useful bioactive compounds?

Also, it is important to point out that, in the experimental laboratory, the "pharmacophore" concept is sometimes rather loosely used to denote whatever structural feature is held responsible for the activity. It is not rare to hear mentioned the "benzodiazepine pharmacophore," in the sense of "benzodiazepine scaffold." For example, Brizzi et al. [107] report "a novel pharmacophore consisting of both a rigid aromatic backbone and a flexible chain with the aim to develop a series of stable and potent ligands of cannabinoid receptors."

The following is a brief and nonexhaustive review of the most recent successful virtual screening applications, followed by experimental validation:

3.1. Recent Structure-Based Success Stories

Type 1 11-β-Hydroxysteroid dehydrogenase (11-β-HSD1) inhibitors were discovered [108] by a Catalyst [109]-based virtual screening, using an active site-derived pharmacophore model generated with LigandScout [110] on the basis of a enzyme-inhibitor complex X-ray structure. Pharmacophore-matching hits were further on filtered by docking into the active site, using GLIDE [111]. Finally, 56 compounds were selected and submitted to biological testing. Eleven compounds with IC_{50} values below 10 μM were found, featuring three new chemical scaffolds as 11β-HSD1 selective inhibitors.

Scaffold hopping (*see* also §3.2) is mostly cited in conjunction to ligand-based design although structure-based design starting from known site–ligand complex structures is as well amenable to the discovery of binders of new chemical classes. In fact, there is no sharp delimitation between the two strategies. The knowledge of the binding geometries from cocrystals X-ray structures may well be used in order to generate a ligand overlay model from the superposition of the entire complexes, thus leaving no more room for doubts concerning the correctness of binding geometries and overlay mode. This ligand overlay may then serve for consensus pharmacophore extraction, its encoding under the form of molecular fingerprints as classically seen in ligand-based approaches, followed by quick database screening. The herein retrieved virtual hits may then be revisited in the structure-based framework and docked into the active site. Such a strategy was recently applied [112] for the discovery of new PPAR-γ agonists, starting from the cocrystals structures of several natural compounds binding the receptor. Common pharmacophore features of considered natural ligands were singled out after the overlay of the corresponding complexes, coded under the form of a LIQUID [113] pharmacophore and used for database screening. Primary hits were docked into the PPAR active site, and two out of eight tested molecules (all chemically different with respect to the initially considered natural products) were found to display significant activity in a cellular reporter gene assay.

3.2. Chemotype-Hopping Applications

Discovery of binders based on radically different chemical structures (chemotype-, scaffold-, or lead-hopping) is the main purpose of pharmacophore-based modeling, the more so knowing that a medicinal chemist's brain may well understand scaffold-based similarity but fails to visualize spatial complementarity of equivalent functional groups. Therefore, the computer is, in this respect, a truly complementary tool to "chemical intuition."

A classical Catalyst [109]-based pharmacophore screening for serotonin 2C receptor ligands, leading to the discovery of novel nanomolar binders, not implying any special tuning or novel computational tool development, was recently reported [114]. It shows that commercial software may provide, as far as pharmacophore searching goes, valid "keys in hands" solutions for the medicinal chemistry lab.

A radical and successful example of discovery of nonsaccharide organic (tertraisoquinoleine-based) molecules to mimic the activating effect by the extremely complex sugar Heparin of the coagulation factor Antithrombin (a plasma glycoprotein serine protease inhibitor) was achieved [115] by means of simple, interactive design of compounds fitting a heparin-inspired pharmacophore.

The discovery [116] of a highly active Carbonic Anhydrase inhibitor, out of as few as six (scaffold-wise unrelated) selected virtual hits, has been achieved by means of a combined strategy involving a cascade of successive ligand-based (ligand overlay-based pharmacophore screening), structure-based (docking into a homology model of the protein), and chemical intuition-driven selection pharmacophore model building. The MOE [117] software suite has been used.

One of the most radical examples of scaffold hopping is the design of dual site inhibitors of macromolecules having a cofactor binding site not far from the actual substrate fixation site. It may then be possible, using the pharmacophore model derived from the ternary protein–substrate–cofactor complex, to virtually screen for compounds simultaneously binding to both sites, which may be entropically beneficial. Of course, such inhibitors will be chemically different from either of the previously known substrate-competitive or cofactor-competitive inhibitors. A recent example [118] targeting Dihydrofolate Reductase (DHFR) of the anthrax bacillus used the Sybyl [119] software suite for pharmacophore screening followed by docking to discover micromolar, allegedly "bidentate" ligands blocking both the methotrexate and the NADPH cofactor site. Out of 15 selected molecules, two displayed low micromolar activity against DHFR. Neither are derivatives of traditional antifolates, an advantage being that these structurally and chemically distinct compounds possibly represent the first leads of two new classes of DHFR inhibitors. It is thus possible to combine key interaction points from different sites into a novel pharmacophore model – if these are reasonably close in space, and the resulting model is not overwhelmingly complex.

Pharmacophores can be successfully applied to model biological effects beyond simple ligand binding, such as, for example, protein–protein heterodimer disruption. Without knowing the exact interaction mechanisms, a ligand-based approach [120], using

GALAHAD [121] allowed to extract common pharmacophore features of a set of known disruptors of the c-Myc-Max dimer, then screen the Zinc database using Sybyl [119]. Nine compounds, none within the chemical class of the pharmacophore training molecules, were tested with significant degree of success in both in vitro and cellular tests.

3.3. There is No Single Best Virtual Screening Approach: The Importance of Testing Alternative Methods and Consensus Scoring

An interesting study [122], leading to the in silico discovery of a potent Human Immunodeficiency Virus (HIV) Entry blocker, binder of the CXCR4 chemokine receptor, outlined the importance of benchmarking a maximum of possible virtual screening tools with respect to the considered target, in order to eventually pick hits among molecules consensually predicted to be active by the methods best performing in the tests. Considered methodologies included both ligand-based approaches (QSAR and pharmacophore modeling with MOE [117] and Discovery Studio [123], shape matching tools like PARAFIT [124], ROCS [125], and HEX) and structure-based approaches (docking with AUTODOCK [126], GOLD [127], FRED [128], and HEX [129]) based on a homology model of the receptor. The methods were fine-tuned in a retrospective virtual screening, and then successfully used in a prospective search.

New Hormone Sensitive Lipase inhibitors were discovered [130] by an original virtual screening approach using QSAR models combining pharmacophore hypothesis matching scores (determined with respect to a basis set of Catalyst [109]-generated pharmacophore hypotheses) and classical molecular descriptors. While hypothesis matching scores, per se, only loosely correlate with observed affinities, they appear to be useful molecular descriptors in QSAR equations. The same strategy was also used for discovery of novel Neuraminidase inhibitors [131].

3.4. Addressing Novel Binding Sites: Design of Noncompetitive Inhibitors

The NS5B protein, an RNA-dependent RNA polymerase, a key target for therapeutic intervention against the Hepatitis C virus, was typically blocked by means of designed nucleoside analogs or mimics, binding at the nucleoside binding site. Nucleosides being a major player of the cellular clockwork – therefore, close structural mimics are at high risk of being bound by other receptors and enzymes, leading to potentially serious side effects. A potential allosteric site of NS5B, distinct from the catalytic center, was targeted by means of structure-based design [132]. By virtual screening, the compound library was down-sized from 3.5 million to 119 chemicals. The inhibitory activities of the selected compounds were tested in vitro and confirmed the discovery of low-potency, but interesting noncompetitive NS5B inhibitors.

3.5. Multi-target Pharmacophore Modeling

The ability to choose interaction features to be included in, respectively left out of, a considered pharmacophore model ensures that the same methodology may be successfully used either for designing specific inhibitors of a target, not hitting related macromolecules – by inclusion of site-specific features – or, rather oppositely, for the design of promiscuous binders expected to hit a large panel of targets of a same biological class. In the latter case, the pharmacophore model should be based only on features that are conserved in all the targets of that class: G-coupled protein receptors (GPCRs), kinases, etc. This can be achieved by means of machine learning [133, 134] from a training set of compounds classified in "binders" and "non-binders" of a target family. How many representatives of the target class have to be "hit" by a molecule, and how strong the interaction has to be in order to have this labeled as representative binder of the entire "class," are matters of empirical choice – and so is the definition of the "target class" (all GPCRs? Rhodopsin-like GPCRs only? etc.) The design of class-specific rather than target-specific inhibitors may be a useful compound library design technique if the library is intended for multiple testing in various bioassays within that class – primary, and likely promiscuous hits having to undergo a specificity-enhancement optimization phase in order to be rendered target-specific. Sometimes, the targeted in vivo activity does not require an absolute specificity with respect to a given target, as in the case of Central Nervous System (CNS) drugs, known to hit multiple GPCRs in the brain, having a therapeutic effect (and, quite often, important side effects) emerging from the subtle interplay of all these interactions. Therefore, in the specific case of CNS activity, it does make sense to go beyond target class-directed library design, to in vivo-effect directed library design. This [135] amounts to searching for a "pharmacophore" guaranteeing a strong affinity for various GPCRs *plus* an overall pharmacophore pattern with a balanced occurrence of polar and hydrophobic groups, in order to ensure the required pharmacokinetic properties of CNS drugs (passage of the blood–brain barrier, in particular).

If biological rationale suggests that simultaneous inhibition of two or more macromolecules is desirable for curing a given disease, and these targets are structurally related, in the sense of conserved key ligand anchoring points, then the design of specific multi-target drugs can be addressed by means of the largest common pharmacophore encoding the conserved anchoring points. Targeting few specific targets rather than an entire class of related proteins has the advantage of allowing individual pharmacophore extraction, comparison, and manual pruning to a subset of common key interactions. Docking calculations of pharmacophore matching compounds can be a valuable filter of primary virtual

hits. A strategy based on the above outlined principles was successfully [136] applied to screen for dual-target inhibitors against both the human leukotriene A4 hydrolase (LTA4H-h) and the human nonpancreatic secretory phospholipase A2 (hnps-PLA2). Three compounds screened from the chemical database MDL Available Chemical Directory were found to inhibit these two enzymes at the 10 μM level.

3.6. Mechanistically Relevant Pharmacophores from Molecular Simulations

An original strategy [137] to block a parasite-specific metabolic pathway in *Plasmodium falciparum* by means of disruption of the bioactive homodimeric form of a specific kinase (CMK) was based on the construction of a pharmacophore based on the protein–protein (direct or water-mediated) contacts that are thought to be mechanistically relevant for dimerization, according to a molecular dynamics study. This pharmacophore model was used for classical database screening. Using an intensity-fading matrix-assisted laser desorption/ionization time-of-flight mass spectrometry approach, one of the virtual hits was found to interact with CMK. This approach suggests that the empirical pharmacophore search tool can be meaningfully used in complement to rigorous molecular simulations, to provide a convenient, synthetic wrap-up of therein extracted, mechanistically relevant information, and to use it for database screening – something that molecular dynamics *per se* is not able to do.

4. Conclusions

In light of the wealth of recent, both methodological and applicative, publications devoted to, or based on, pharmacophore modeling, it can be safely concluded that these techniques form, next to (sub)structure-based queries, the backbone of modern chemoinformatics in modern drug design. Reducing the complexity of non-covalent interactions to a set of rules describing the behavior of atom groups characterized by their pharmacophore types turned out to be a fruitful idea, leading to an approach that is both simple enough to be understood and accepted by bench chemists and realistic enough to allow for verifiable and successful predictions. This domain has doubtlessly reached maturity – a plethora of commercial and free software suites supports pharmacophore modeling in all its variants – topological or three-dimensional, structure and ligand-based, overlay-based or overlay-free.

However, on one hand, after browsing through the latest methodological papers in the field, it may be argued that there is still room for technological improvement – in terms of better conformational sampling, more rigorous pharmacophore typing schemes, faster flexible overlay techniques, intelligent selection of

potential anchoring points in protein sites, etc. Yet, at second thought, it is not clear at all whether further methodological progress will significantly enhance the performance of pharmacophore-based virtual screening. Pharmacophore modeling is not a fundamental theory of matter, and, as such, intrinsically error-prone. Furthermore, as a series of successive modeling steps – say, in typical ligand-based procedures, conformational sampling, followed by pharmacophore typing, followed by molecular overlay, followed by hypothesis extraction and eventually hypothesis scoring – its overall performance cannot be better than the one supported by the weakest link of this chain. In this sense, specific work to improve one or the other of these aspects may not lead to overall benefits. More fine-grained conformational sampling may reduce the chance to miss the "bioactive" conformer, but may as well "drown" it in a large pool of irrelevant geometries. In ligand-based approaches, conformational sampling will be intrinsically flawed because of the impossibility to prioritize the bioactive conformer over the others when its binding site geometry is not known. The hypothesis that flexible multiple overlays of known actives must forcibly lead to the discovery of bioactive conformers (the ones that stand out as the only geometries that are simultaneously compatible with a meaningful overlay) is not much more than wishful thinking when more than 3 rotatable bonds/ligand are involved. On one hand, the volume of the problem space exponentially explodes as a function of the number of ligands; and on the other, the empirical question of how much strain energy/ligand can be tolerated in order to improve the overlay quality score is incontrovertible and does not accept any physically meaningful answers. Next, better conformational sampling may strictly make no difference if pharmacophore typing is not realistic. Yet, protonation state predictions – and, notably, pK_a shifts induced by the actual binding to the receptor – are very difficult.

The importance of the methodological novelties of the latest years is therefore very difficult to assess. They may be intellectually sound and appealing, yet it is not at all sure that they will significantly improve hit rates in pharmacophore-based virtual screens. In a larger context, the absence of reports of pharmacophore-based screening *failures* renders the objective estimation of the robustness of various methods impossible. This is a fundamental problem of the entire field of computer-aided molecular simulations in drug design, for failures, even if reported, are difficult to interpret – were they imputable to the methodology or were no actives found because the screened data base did not contain any (at least not any that should have been found, within the applicability domain of the method). It is not easy to apply Occam's razor to "shave" the irrelevant refinement off the latest developments in pharmacophore modeling.

Although purely theoretical pharmacophore constructs and structure–activity relations are still being published *per se*, without any experimental follow-up (and without introducing any methodological novelties), encouragingly, the majority of the reported applications of pharmacophore modeling were sustained by actual experimental validation of the virtual hits. In the end, it is less important to know whether a method is slightly "better" than another, statistically speaking, than it is to know whether a method can provide experimentally valid responses. In this sense, the independent validation of out-of-box technological solutions by independent research groups – commercial software used as such, without any user-added improvements was repeatedly shown to yield valid results – is excellent news. Unfortunately, failures not being reported, an objective cost/benefit analysis of the use of pharmacophore-based virtual screening in industry and academia cannot be undertaken. Furthermore, it should be kept in mind that any rational drug design undertaken instead of a random screening campaign will – likely – be much more cost-efficient but unlikely to lead to any paradigm breaking discoveries in the field of interest. Scaffold hopping is as "revolutionary" as pharmacophore modeling can get: it may find new chemotypes matching new binding modes. As always, rational approaches are here to refine serendipitous discoveries.

References

1. IUPAC. (2007) Glossary of Terms used in Medicinal Chemistry, (IUPAC, Ed.), IUPAC.
2. Jorgensen, W. L. (1991) Rusting of the lock and key model for protein-ligand binding. *Science* **254**, 954–955.
3. Choudhury, N., Montgomery-Pettitt, B. (2007) The dewetting transition and the hydrophobic effect. *Journal of the American Chemical Society* **129**, 4847–4852.
4. Thomson Reuters. (2009) ISI Web of Knowledge, New York.
5. Wang, C., Bradley, P., and Baker, D. (2007) Protein-protein docking with backbone flexibility. *Journal of Molecular Biology* **373**, 503–519.
6. De Grandis, V., Bizzarri, A. R., and Cannistraro, S. (2007) Docking study and free energy simulation of the complex between p53 DNA-binding domain and azurin. *Journal of Molecular Recognition* **20**, 215–226.
7. Guvench, O., and MacKerell, A. D., Jr. (2008) Comparison of protein force fields for molecular dynamics simulations. *Methods in Molecular Biology*, 63–88.
8. Ponder, J. W., and Case, D.A. (2003) Force fields for protein simulations. *Advances in Protein Chemistry* **66**, 27–85.
9. Parent, B., Kökösy, A., and Horvath, D. (2007) Optimized evolutionary strategies in conformational sampling. *Soft Computing* **11**, 63–79.
10. Horvath, D. (2008) Topological Pharmacophores. in *Chemoinformatics Approaches to Virtual Scrrening* (Varnek, A., and Tropsha, A., Eds.), pp 44–72, RCS Publishing, Cambridge, UK.
11. Bergmann, R., Linusson, A., and Zamora, I. (2007) SHOP: Scaffold HOPping by GRID-based similarity searches. *Journal of Medicinal Chemistry* **50**, 2708–2717.
12. Poulain, R., Horvath, D., Bonnet, B., Eckoff, C., Chapelain, B., Bodinier, M-C., and Deprez, B. (2001) From hit to lead. Combining two complementary methods for focused library design application to μ opiate ligands. *Journal of Medicinal Chemistry* **44**, 3378–3390.
13. Schlosser, J., and Rarey, M. (2009) Beyond the virtual screening paradigm: Structure-based searching for new lead compounds.

Journal of Chemical Information and Modeling **49**, 800–809.
14. Koppen, H. (2009) Virtual screening – What does it give us? *Current Opinion in Drug Discovery & Development* **12**, 397–407.
15. Sun, H. M. (2008) Pharmacophore-based virtual screening. *Current Medicinal Chemistry* **15**, 1018–1024.
16. Sperandio, O., Miteva, M. A., and Villoutreix, B. O. (2008) Combining ligand- and structure-based methods in drug design projects. *Current Computer-Aided Drug Design* **4**, 250–258.
17. Prinz, H. (2008) How to identify a pharmacophore. *Chemistry & Biology* **15**, 207–208.
18. Muegge, I. (2008) Synergies of virtual screening approaches. *Mini-Reviews in Medicinal Chemistry* **8**, 927–933.
19. Mauser, H., and Guba, W. (2008) Recent developments in de novo design and scaffold hopping. *Current Opinion in Drug Discovery & Development* **11**, 365–374.
20. Green, D. V. S. (2008) Virtual screening of chemical libraries for drug discovery. *Expert Opinion on Drug Discovery* **3**, 1011–1026.
21. Douguet, D. (2008) Ligand-based approaches in virtual screening. *Current Computer-Aided Drug Design* **4**, 180–190.
22. Van Drie, J. H. (2007) Computer-aided drug design: The next 20 years. *Journal of Computer-Aided Molecular Design* **21**, 591–601.
23. McInnes, C. (2007) Virtual screening strategies in drug discovery. *Current Opinion in Chemical Biology* **11**, 494–502.
24. Mason, J. S., Good, A. C., and Martin, E. J. (2001) 3-D pharmacophores in drug discovery. *Current Pharmaceutical Design* **7**, 567–597.
25. Güner, O. F. (2000) *Pharmacophore Perception, Use and Development in Drug Design*, International University Line, La Jolla, CA.
26. Orts, J., Grimm, S. K., Griesinger, C., Wendt, K. U., Bartoschek, S., and Carlomagno, T. (2008) Specific methyl group protonation for the measurement of pharmacophore-specific interligand NOE interactions. *Chemistry A European Journal* **14**, 7517–7520.
27. Bonachera, F., Parent, B., Barbosa, F., Froloff, N., and Horvath, D. (2006) Fuzzy tricentric pharmacophore fingerprints. 1 – Topological fuzzy pharmacophore triplets and adapted molecular similarity scoring schemes. *Journal of Chemical Information and Modeling* **46**, 2457–2477.
28. Guha, R., and Van Drie, J. H. (2008) Structure-activity landscape index: Identifying and quantifying activity cliffs. *Journal of Chemical Information and Modeling* **48**, 646–658.
29. Maggiora, G. M. (2006) On outliers and activity cliffs – Why QSAR often disappoints. *Journal of Chemical Information and Modeling* **46**, 1535–1535.
30. Bonachera, F., and Horvath, D. (2008) Fuzzy tricentric pharmacophore fingerprints. 2. Application of topological fuzzy pharmacophore triplets in quantitative structure-activity relationships. *Journal of Chemical Information and Modeling* **48**, 409–425.
31. Horvath, D., Mao, B., Gozalbes, R., Barbosa, F., and Rogalski, S. L. (2004) Strengths and Limitations of Pharmacophore-Based Virtual Screening. in *Chemoinformatics in Drug Discovery*. (Oprea, T. I., Ed.), pp 117–137, WILEY-VCH Verlag GmbH, Weinheim.
32. von Korff, M., Freyss, J., and Sander, T. (2008) Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *Journal of Chemical Information and Modeling* **48**, 797–810.
33. von Korff, M., Freyss, J., and Sander, T. (2008) Comparison of Ligand- and Structure-Based Virtual Screening on the DUD Data Set. in *8th International Conference on Chemical Structures*, pp 209–231, Amer Chemical Soc, Noordwijkerhout, Netherlands.
34. Cramer, R. D., Patterson D. E., and Bunce, J. E. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* **110**, 5959–5967.
35. Manallack, D. T. (2008) The use of local surface properties for molecular superimposition. *Journal of Molecular Modeling* **14**, 797–805.
36. Sperandio, O., Souaille, M., Delfaud, F., Miteva, M. A., and Villoutreix, B. O. (2009) MED-3DMC: A new tool to generate 3D conformation ensembles of small molecules with a Monte Carlo sampling of the conformational space. *European Journal of Medicinal Chemistry* **44**, 1405–1409.
37. Liu, X. F., Bai, F., Ouyang, S. S., Wang, X. C., Li, H. L., and Jiang, H. L. (2009) Cyndi: A multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics* **10**, 14.
38. Li, J., Ehlers, T., Sutter, J., Varma-O'Brien, S., and Kirchmair, J. (2007) CAESAR: A new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *Journal of*

Chemical Information and Modeling **47**, 1923–1932.
39. Takagi, T., Amano, M., and Tomimoto, M. (2009) Novel method for the evaluation of 3D conformation generators. *Journal of Chemical Information and Modeling* **49**, 1377–1388.
40. Perola, E., and Charifson, P. S. (2004) Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *Journal of Medicinal Chemistry* **47**, 2499–2510.
41. Böhm, H. J. (1992) The Computer Program LUDI: A new method for the de novo design of enzyme inhibitors. *Journal of Computer-Aided Molecular Design* **6**, 61–78.
42. Gillet, V., Johnson, A. P., Mata, P., Sike, S., and Williams, P. (1993) SPROUT: A program for structure generation. *Journal of Computer-Aided Molecular Design* **7**, 127–153.
43. Murray, C. W., Clark, D. E., Auton, T. R., Firth, M. A., Li, J., Sykes, R. A., Waszkowycz, B., Westhead, D. R., and Young, S. C. (1997) PRO_SELECT: Combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *Journal of Computer-Aided Molecular Design* **11**, 193–207.
44. Tintori, C., Corradi, V., Magnani, M., Manetti, F., and Botta, M. (2008) Targets looking for drugs: A multistep computational protocol for the development of structure-based pharmacophores and their applications for hit discovery. *Journal of Chemical Information and Modeling* **48**, 2166–2179.
45. Goodford, P. J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* **28**, 849–857.
46. Bowman, A. L., Lerner, M. G., and Carlson, H. A. (2007) Protein flexibility and species specificity in structure-based drug discovery: Dihydrofolate reductase as a test system. *Journal of the American Chemical Society* **129**, 3634–3640.
47. Damm, K. L., and Carlson, H. A. (2007) Exploring experimental sources of multiple protein conformations in structure-based drug design. *Journal of the American Chemical Society* **129**, 8225–8235.
48. Jayachandran, G., Shirts, M. R., Park, S., and Pande, V. S. (2006) Parallelized over parts computation of absolute binding free energy with docking and molecular dynamics. *The Journal of Chemical Physics* **125**, 84901–84905.
49. Barillari, C., Marcou, G., and Rognan, D. (2008) Hot-spots-guided receptor-based pharmacophores (HS-Pharm): A knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *Journal of Chemical Information and Modeling* **48**, 1396–1410.
50. Martin, E. J., and Sullivan, D. C. (2008) Surrogate AutoShim: Predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure. *Journal of Chemical Information and Modeling* **48**, 873–881.
51. Zou, J., Xie, H. Z., Yang, S. Y., Chen, J. J., Ren, J. X., and Wei, Y. Q. (2008) Towards more accurate pharmacophore modeling: Multicomplex-based comprehensive pharmacophore map and most-frequent-feature pharmacophore model of CDK2. *Journal of Molecular Graphics* **27**, 430–438.
52. Ebalunode, J. O., Ouyang, Z., Liang, J., and Zheng, W. (2008) Novel approach to structure-based pharmacophore search using computational geometry and shape matching techniques. *Journal of Chemical Information and Modeling* **48**, 889–901.
53. Horvath, D. (2001) ComPharm – Automated comparative analysis of pharmacophoric patterns and derived QSAR approaches, novel tools in high throughput drug discovery. A proof of concept study applied to farnesyl protein transferase inhibitor design. in *QSPR/QSAR Studies by Molecular Descriptors* (Diudea, M. V., Ed.), pp 395–439., Nova Science Publishers, Inc, New York.
54. Landrum, G. A., Penzotti, J. E., and Putta, S. (2006) Feature-map vectors: A new class of informative descriptors for computational drug discovery. *Journal of Computer-Aided Molecular Design* **20**, 751–762.
55. Totrov, M. (2008) Atomic property fields: Generalized 3D pharmacophoric potential for automated ligand superposition, pharmacophore elucidation and 3D QSAR. *Chemical Biology & Drug Design* **71**, 15–27.
56. Putta, S., Landrum, G. A., and Penzotti, J. E. (2005) Conformation mining: An algorithm for finding biologically relevant conformations. *Journal of Medicinal Chemistry* **48**, 3313–3318.
57. Taminau, J., Thijs, G., and De Winter, H. (2008) Pharao: Pharmacophore alignment and optimization. *Journal of Molecular Graphics & Modelling* **27**, 161–169.
58. Todorov, N. P., Alberts, I. L., de Esch, I. J. P., and Dean, P. M. (2007) QUASI: A

novel method for simultaneous superposition of multiple flexible ligands and virtual screening using partial similarity. *Journal of Chemical Information and Modeling* **47**, 1007–1020.
59. Wolber, G., Dornhofer, A. A., and Langer, T. (2006) Efficient overlay of small organic molecules using 3D pharmacophores. *Journal of Computer-Aided Molecular Design* **20**, 773–788.
60. Bandyopadhyay, D., and Agrafiotis, D. K. (2008) A self-organizing algorithm for molecular alignment and pharmacophore development. *Journal of Computational Chemistry* **29**, 965–982.
61. Cottrell, S. J., Gillet, V. J., and Taylor, R. (2006) Incorporating partial matches within multiobjective pharmacophore identification. *Journal of Computer-Aided Molecular Design* **20**, 735–749.
62. Nandigam, R. K., Evans, D. A., Erickson, J. A., Kim, S., and Sutherland, J. J. (2008) Predicting the Accuracy of ligand overlay methods with random forest models. *Journal of Chemical Information and Modeling* **48**, 2386–2394.
63. Hähnke, V., Hofmann, B., Grgat, T., Proschak, E., Steinhilber, D., and Schneider, G. (2009) PhAST: Pharmacophore alignment search tool. *Journal of Computational Chemistry* **30**, 761–771.
64. Rafael Gozalbes, F. B., Nicolaï, E., Horvath, D., Froloff, N. (2009) Development and validation of a pharmacophore-based QSAR model for the prediction of CNS activity. *ChemMedChem* **4**, 204–209.
65. Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., Labaudiniere, R. F. (1998) New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry* **38**, 144–150.
66. Shepphird, J. K., and Clark, R. D. (2006) A marriage made in torsional space: Using GALAHAD models to drive pharmacophore multiplet searches. *Journal of Computer-Aided Molecular Design* **20**, 763–771.
67. Sciabola, S., Morao, I., and de Groot, M. J. (2007) Pharmacophoric fingerprint method (TOPP) for 3D-QSAR modeling: Application to CYP2D6 metabolic stability. *Journal of Chemical Information and Modeling* **47**, 76–84.
68. Watson, P. (2008) Naive Bayes classification using 2D pharmacophore feature triplet vectors. *Journal of Chemical Information and Modeling* **48**, 166–178.
69. Askjaer, S., and Langgard, M. (2008) Combining pharmacophore fingerprints and PLS-discriminant analysis for virtual screening and SAR elucidation. *Journal of Chemical Information and Modeling* **48**, 476–488.
70. Podolyan, Y., and Karypis, G. (2009) Common pharmacophore identification using frequent clique detection algorithm. *Journal of Chemical Information and Modeling* **49**, 13–21.
71. Sperandio, O., Andrieu, O., Miteva, M. A., Vo, M. Q., Souaille, M., Delfaud, F., and Villoutreix, B. O. (2007) MED-SuMoLig: A new ligand-based screening tool for efficient scaffold hopping. *Journal of Chemical Information and Modeling* **47**, 1097–1110.
72. Buttingsrud, B., King, R. D., and Alsberg, B. K. (2007) An alignment-free methodology for modelling field-based 3D-structure activity relationships using inductive logic programming. *Journal of Chemometrics* **21**, 509–519.
73. Tsunoyama, K., Amini, A., Sternberg, M. J. E., and Muggleton, S. H. (2008) Scaffold hopping in drug discovery using inductive logic programming. *Journal of Chemical Information and Modeling* **48**, 949–957.
74. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-Hopping" by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie* **38**, 2894–2896.
75. Zhu, F. Q., and Agrafiotis, D. K. (2007) Recursive distance partitioning algorithm for common pharmacophore identification. *Journal of Chemical Information and Modeling* **47**, 1619–1625.
76. Kirchmair, J., Ristic, S., Eder, K., Markt, P., Wolber, G., Laggner, C., and Langer, T. (2007) Fast and efficient in silico 3D screening: Toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *Journal of Chemical Information and Modeling* **47**, 2182–2196.
77. Güner, O., Clement, O., and Kurogi, Y. (2004) Pharmacophore modeling and three dimensional database searching for drug design using catalyst: Recent advances. *Current Medicinal Chemistry* **11**, 2991–3005.
78. Kurogi, Y., and Güner, O. (2001) Pharmacophore modeling and threedimensional database searching for drug design using catalyst. *Current Medicinal Chemistry* **8**, 1035–1055.
79. Matter, H., and Pötter, T. (1999) Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound

subsets. *Journal of Chemical Information and Modeling* **39**, 1211–1225.
80. Horvath, D., and Jeandenans, C. (2003) Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces – A benchmark for neighborhood behavior assessment of different in silico similarity metrics. *Journal of Chemical Information and Computer Sciences* **43**, 691–698.
81. Fox, P. C., Wolohan, P. R. N., Abrahamian, E., and Clark, R. D. (2008) Parameterization and conformational sampling effects in pharmacophore multiplet searching. *Journal of Chemical Information and Modeling* **48**, 2326–2334.
82. Nisius, B., Vogt, M., and Bajorath, J. (2009) Development of a fingerprint reduction approach for Bayesian similarity searching based on Kullback-Leibler divergence analysis. *Journal of Chemical Information and Modeling* **49**, 1347–1358.
83. Kirchmair, J., Wolber, G., Laggner, C., and Langer, T. (2006) Comparative performance assessment of the conformational model generators omega and catalyst: A large-scale survey on the retrieval of protein-bound ligand conformations. *Journal of Chemical Information and Modeling* **46**, 1848–1861.
84. Kirchmair, J., Laggner, C., Wolber, G., and Langer, T. (2005) Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *Journal of Chemical Information and Modeling* **45**, 422–430.
85. Chen, I. J., and Foloppe, N. (2008) Conformational sampling of druglike molecules with MOE and catalyst: Implications for pharmacophore modeling and virtual screening. *Journal of Chemical Information and Modeling* **48**, 1773–1791.
86. Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409–443.
87. Jewsbury, P. J., Taylor, R. D., and Essex, J. W. (2002) A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design* **16**, 151–166.
88. Rarey, M., Claussen, H., Buning, C., and Lengauer, T. (2001) FlexE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology* **308**, 377–395.
89. Todd, J., Ewing, A., and Kuntz, I. D. (1998) Critical evaluation of search algorithms for automated molecular docking and database screening. *Journal of Computational Chemistry* **18**, 1175–1189.
90. Muthas, D., Sabnis, Y. A., Lundborg, M., and Karlen, A. (2008) Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering. *Journal of Molecular Graphics* **26**, 1237–1251.
91. Brewerton, S. C. (2008) The use of protein-ligand interaction fingerprints in docking. *Current Opinion in Drug Discovery & Development* **11**, 356–364.
92. Venhorst, J., Nunez, S., Terpstra, J. W., and Kruse, C. G. (2008) Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *Journal of Medicinal Chemistry* **51**, 3222–3229.
93. Marcou, G., and Rognan, D. (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling* **47**, 195–207.
94. Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., and Mason, J. S. (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application. *Journal of Chemical Information and Modeling* **47**, 279–294.
95. Horvath, D., Marcou, G., and Varnek, A. (2009) Predicting the predictability: A unified approach to the applicability domain problem of QSAR models. *Journal of Chemical Information and Modeling* **49**, 1762–1776.
96. Tetko, I. V., Sushko, I., Pandey, A. K., Zhu, H., Tropsha, A., Papa, E., Oberg, T., Todeschini, R., Fourches, D., and Varnek, A. (2008) Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: Focusing on applicability domain and overfitting by variable selection. *Journal of Chemical Information and Modeling* **48**, 1733–1746.
97. Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M., McDowell, R. M., and Gramatica, P. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *QSAR & Combinatorial Science* **111**, 1361–1375.
98. Ma, X. H., Jia, J., Zhu, F., Xue, Y., Li, Z. R., and Chen, Y. Z. (2009) Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Combinatorial Chemistry & High Throughput Screening* **12**, 344–357.

99. Klebe, G. (2008) Understanding QSAR: Do we always use the correct structural models to establish affinity correlation?
100. Gonzalez, M. P., Teran, C., Saiz-Urra, L., and Teijeira, M. (2008) Variable selection methods in QSAR: An overview. *Current Topics in Medicinal Chemistry* **8**, 1606–1627.
101. Nair, P. C., and Sobhia, M. E. (2008) Fingerprint directed scaffold hopping for identification of CCR2 antagonists. *Journal of Chemical Information and Modeling* **48**, 1891–1902.
102. Mascarenhas, N. M., and Ghoshal, N. (2008) An efficient tool for identifying inhibitors based on 3D-QSAR and docking using feature-shape pharmacophore of biologically active conformation – A case study with CDK2/CyclinA. *European Journal of Medicinal Chemistry* **43**, 2807–2818.
103. Vadivelan, S., Sinha, B. N., Tajne, S., and Jagarlapudi, S. (2009) Fragment and knowledge-based design of selective GSK-3 beta inhibitors using virtual screening models. *European Journal of Medicinal Chemistry* **44**, 2361–2371.
104. Dong, A. G., Huo, J. F., Gao, Q. Z., Zhao, K., and Wei, J. (2009) A three-dimensional pharmacophore model for RXR alpha agonists. *Journal of Molecular Structure* **920**, 252–263.
105. Xie, Q. Q., Xie, H. Z., Ren, J. X., Li, L. L., and Yang, S. Y. (2009) Pharmacophore modeling studies of type I and type II kinase inhibitors of Tie2. *Journal of Molecular Graphics* **27**, 751–758.
106. Andrade, C. H., Pasqualoto, K. F. M., Ferreira, E. I., and Hopfinger, A. J. (2009) Rational design and 3D-pharmacophore mapping of 5′-thiourea-substituted alpha-thymidine analogues as mycobacterial TMPK inhibitors. *Journal of Chemical Information and Modeling* **49**, 1070–1078.
107. Brizzi, A., Brizzi, V., Cascio, M. G., Corelli, F., Guida, F., Ligresti, A., Maione, S., Martinelli, A., Pasquini, S., Tuccinardi, T., and Di Marzo, V. (2009) New resorcinol-anandamide "hybrids" as potent cannabinoid receptor ligands endowed with antinociceptive activity in vivo. *Journal of Medicinal Chemistry* **52**, 2506-2514.
108. Yang, H. Y., Shen, Y., Chen, J. H., Jiang, Q. F., Leng, Y., and Shen, J. H. (2009) Structure-based virtual screening for identification of novel 11 beta-HSD1 inhibitors. *European Journal of Medicinal Chemistry* **44**, 1167–1171.
109. Accelrys Software, I. (2006) Catalyst, 4.9 ed., San Diego.
110. Wolber, G., and Langer, T. (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *Journal of Chemical Information and Modeling* **45**, 160–169.
111. Schrödinger, L. (2005) Glide, New York.
112. Tanrikulu, Y., Rau, O., Schwarz, O., Proschak, E., Siems, K., Muller-Kuhrt, L., Schubert-Zsilavecz, M., and Schneider, G. (2009) Structure-based pharmacophore screening for natural-product-derived PPAR gamma agonists. *Chembiochem* **10**, 75–78.
113. Tanrikulu, Y., Nietert, M., Scheffer, U., Proschak, E., Grabowski, K., Schneider, P., Weidlich, M., Karas, M., Goebel, M., and Schneider, G. (2007) Scaffold hopping by "fuzzy" pharmacophores and its application to RNA targets. *Chembiochem* **8**, 1932–1936.
114. Ahmed, A., Choo, H., Cho, Y. S., Park, W. K., and Pae, A. N. (2009) Identification of novel serotonin 2C receptor ligands by sequential virtual screening. *Bioorganic & Medicinal Chemistry* **17**, 4559–4568.
115. Raghuraman, A., Liang, A. Y., Krishnasamy, C., Lauck, T., Gunnarsson, G. T., and Desai, U. R. (2009) On designing non-saccharide, allosteric activators of antithrombin. *European Journal of Medicinal Chemistry* **44**, 2626–2631.
116. Thiry, A., Ledecq, M., Cecchi, A., Frederick, R., Dogne, J. M., Supuran, C. T., Wouters, J., and Masereel, B. (2009) Ligand-based and structure-based virtual screening to identify carbonic anhydrase IX inhibitors. *Bioorganic & Medicinal Chemistry* **17**, 553–557.
117. (2005) MOE (Molecular Operating Environment), 2005.06 ed., Chemical Computing Group, Inc., Montreal.
118. Bennett, B. C., Wan, Q., Ahmad, M. F., Langan, P., and Dealwis, C. G. (2009) X-ray structure of the ternary MTX.NADPH complex of the anthrax dihydrofolate reductase: A pharmacophore for dual-site inhibitor design. *Journal of Structural Biology* **166**, 162–171.
119. Tripos, I. (2007) Sybyl, 8.0 ed., St. Louis, MO.
120. Mustata, G., Follis, A. V., Hammoudeh, D. I., Metallo, S. J., Wang, H. B., Prochownik, E. V., Lazo, J. S., and Bahar, I. (2009) Discovery of novel Myc-Max heterodimer disruptors with a three-dimensional pharmacophore model. *Journal of Medicinal Chemistry* **52**, 1247–1250.
121. Richmond, N. J., Abrams, C. A., Wolohan, P. R. N., Abrahamian, E., Willett, P., and Clark, R. D. (2006) GALAHAD: 1. Pharmacophore identification by

hypermolecular alignment of ligands in 3D. *Journal of Computer-Aided Molecular Design* **20**, 567–587.
122. Perez-Nueno, V. I., Pettersson, S., Ritchie, D. W., Borrell, J. I., and Teixido, J. (2009) Discovery of novel HIV entry inhibitors for the CXCR4 receptor by prospective virtual screening. *Journal of Chemical Information and Modeling* **49**, 810–823.
123. Accelrys Software, I. (2007) Discovery Studio, 2.0 ed., San Diego, CA.
124. Lin, J., and Clark, T. (2005) An analytical, variable resolution, complete description of static molecules and their intermolecular binding properties. *Journal of Chemical Information and Modeling* **45**, 1010–1016.
125. Grant, A. J., and Pickup, B. T. (1996) A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *Journal of Computational Chemistry* **17**, 1653–1659.
126. Morris, G. M. (2007) AutoDock.
127. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003) Improved protein-ligand docking using GOLD. *Proteins* **52**, 609–623.
128. McGann, M. R., Almond, H. R., Nicholls, A., Grant, J. A., and Brown, F. K. (2003) Gaussian docking functions. *Biopolymers* **68**, 76–90.
129. Ritchie, D. W., and Kemp, G. J. L. (2000) Protein docking using spherical polar Fourier correlations. *Proteins* **39**, 178–194.
130. Taha, M. O., Dahabiyeh, L. A., Bustanji, Y., Zalloum, H., and Saleh, S. (2008) Combining ligand-based pharmacophore modeling, quantitative structure-activity relationship analysis and in silico screening for the discovery of new potent hormone sensitive lipase inhibitors. *Journal of Medicinal Chemistry* **51**, 6478–6494.
131. Abu Hammad, A. M., and Taha, M. O. (2009) Pharmacophore modeling, quantitative structure-activity relationship analysis, and shape-complemented in silico screening allow access to novel influenza neuraminidase inhibitors. *Journal of Chemical Information and Modeling* **49**, 978–996.
132. Ryu, K., Kim, N. D., Choi, S. I., Han, C. K., Yoon, J. H., No, K. T., Kim, K. H., and Seong, B. L. (2009) Identification of novel inhibitors of HCV RNA-dependent RNA polymerase by pharmacophore-based virtual screening and in vitro evaluation. *Bioorganic & Medicinal Chemistry* **17**, 2975–2982.
133. Rolland, C., Gozalbes, R., Nicolai, E., Paugam, M. F., Coussy, L., Barbosa, F., Horvath, D., and Revah, F. (2005) G-protein-coupled receptor affinity prediction based on the use of a profiling dataset: QSAR design, synthesis, and experimental validation. *Journal of Medicinal Chemistry* **48**, 6563–6574.
134. Gozalbes, R., Rolland C., Nicolaï, E., Paugam M.-F., Coussy L., Horvath D., Barbosa F., Mao B., Revah F., and Froloff, N. (2005) QSAR strategy and experimental validation for the development of a GPCR focused library. *QSAR & Combinatorial Science* **24**, 508–516.
135. Gozalbes, R., Barbosa, F., Nicolai, E., Horvath, D., and Froloff, N. (2009) Development and validation of a pharmacophore-based QSAR model for the prediction of CNS activity. *ChemMedChem* **4**, 204–209.
136. Wei, D. G., Jiang, X. L., Zhou, L., Chen, J., Chen, Z., He, C., Yang, K., Liu, Y., Pei, J. F., and Lai, L. H. (2008) Discovery of multitarget inhibitors by combining molecular docking with common pharmacophore matching. *Journal of Medicinal Chemistry* **51**, 7882–7888.
137. Gimenez-Oya, V., Villacanas, O., Fernandez-Busquets, X., Rubio-Martinez, J., and Imperial, S. (2009) Mimicking direct protein-protein and solvent-mediated interactions in the CDP-methylerythritol kinase homodimer: A pharmacophore-directed virtual screening approach. *Journal of Molecular Modeling* **15**, 997–1007.

Chapter 12

De Novo Drug Design

Markus Hartenfeller and Gisbert Schneider

Abstract

Computer-assisted molecular design supports drug discovery by suggesting novel chemotypes and compound modifications for lead structure optimization. While the aspect of synthetic feasibility of the automatically designed compounds has been neglected for a long time, we are currently witnessing an increased interest in this topic. Here, we review state-of-the-art software for de novo drug design with a special emphasis on fragment-based techniques that generate druglike, synthetically accessible compounds. The importance of scoring functions that can be used to predict compound reactivity and potency is highlighted, and several promising solutions are discussed. Recent practical validation studies are presented that have already demonstrated that rule-based fragment assembly can result in novel synthesizable compounds with druglike properties and a desired biological activity.

Key words: Molecular design, Synthesis, Drug-likeness, Reaction, Optimization, Lead structure

1. Introduction

Automated de novo drug design was introduced approximately 20 years ago [1–3]. It has contributed to drug discovery projects ever since by suggesting novel molecular structures with desired properties from scratch and has become a most active field of research during the past few years (Table 1). In an attempt to design innovative bioactive compounds, a medicinal chemist – and equally de novo molecule design software – is confronted with a virtually infinite search space. Search algorithms either come up with an approximate solution, e.g., by stochastic sampling, or restrict the search to a defined section of chemical space which can be screened exhaustively [19, 20]. It is also common to differentiate between deterministic and nondeterministic methods. For example, multiple runs with de novo design software will generate different compounds when a stochastic search algorithm is used. Irrespective of the algorithmic nature of a de novo

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_12, © Springer Science+Business Media, LLC 2011

Table 1
Selected de novo design software approaches published since 2005. For a comprehensive review of earlier methods, see [2]

Software name or author	Year of publication	Required input	
		Reference ligand (ligand-based)	Receptor structure (receptor-based)
Nikitin et al. [4]	2005		X
LEA3D [5]	2005		X
Flux [6, 7]	2006/07	X	
FlexNovo [8]	2006		X
Feher et al. [9]	2008	X	
GANDI [10]	2008	X	X
SQUIRRELnovo [11, 12]	2009	X	
Hecht & Fogel [13]	2009	X	X
FOG [14]	2009	X	
MED-Hybridise [15]	2009		X
MEGA [16]	2009	X	X
Fragment Shuffling [17]	2009	X	X
AutoGrow [18]	2009		X

design software, it is important to keep in mind that it is impossible to consider all theoretically possible virtual molecules due to "combinatorial explosion" [14, 21]. One may well ask the question how automated de novo design can be successful at all. We will present solutions to this problem in this chapter. The trick is to incorporate as much chemical knowledge as possible about the structure of the search space into the design algorithm. *Positive design* restricts the search to regions of chemical space that have a higher probability to find candidate molecules with desired properties. *Negative design*, in contrast, defines criteria that help prevent adverse properties and unwanted chemical structures [22, 23].

It is also fair to say that virtual compound construction software attempts to mimic a synthetic chemist while scoring functions perform virtual assays [24]. The evaluation of candidate compounds plays a critical role in the design process as a de novo design program usually suggests a large number of candidate compounds. Scoring functions indicate which structures are the most promising ones. In this context, it is common to differentiate between *receptor-based* (e.g., docking, receptor-derived

pharmacophores) and *ligand-based* (e.g., similarity metrics) scoring, depending on the reference knowledge used to guide the search for new compounds (Table 1) [19]. Multiple scoring functions can be used in parallel to enable *multiobjective* design [25, 26], i.e., different (potentially competing) properties considered at the same time. For example, properties like aqueous solubility, toxic characteristics, synthetical feasibility, and biological activity are of crucial interest for potential ligand structures and can be explicitly incorporated into the construction process by multiobjective scoring.

In the ideal case, de novo design software suggests high-quality (with regard to the scoring function) molecular structures. Still, there is no guarantee that a designed compound will find the immediate appraisal from a medicinal chemist. It is essential to understand that de novo design rarely yields new chemotypes with nanomolar activity, target selectivity, and an acceptable pharmacokinetic profile. Instead, de novo generated molecules often represent "concept compounds" that require significant further optimization. What can be expected from de novo design, however, is an increased hit rate compared to the screening of an arbitrary compound collection.

For successful automated compound design, three problems must be solved by a de novo design algorithm [2]:

1. *The structure sampling problem* – how to assemble candidate compounds, e.g., atom-based or fragment-based.
2. *The scoring problem* – how to evaluate molecule quality, e.g., by 3D receptor-ligand docking and scoring (requires receptor structure), or ligand-based similarity measure (requires reference ligands, also termed "templates").
3. *The optimization problem* – how to systematically navigate in search space, e.g., by depth-first/breadth-first search, Monte Carlo sampling with Metropolis criterion, evolutionary algorithms, or exhaustive structure enumeration.

Most of the early de novo design tools were strictly atom-based. Modern approaches often provide a diverse selection of large and small virtual molecular entities for compound construction including a few single-atom fragments. Atom-based approaches have the advantages that fine-grained molecule sculpting can be performed and – though only theoretically – the complete chemical universe of structures could be assembled. These advantages come at a price: the huge number of potential solutions complicates a systematic search for actually useful compounds. A shortcut to generating new ligands is the fragment-based approach, by which the size of the search space can be reduced significantly. If fragments are used for molecule assembly that commonly occur in drug molecules, the designed compounds

have a high chance of being druglike themselves. Notably, fragment hits typically possess high "ligand efficiency" [27, 28] (calculated as binding energy divided by the number of heavy atoms) rendering them well suited for further optimization. A fragment can be anything from a single atom to a polycyclic ring system.

For de novo designed candidate ligands, it is insufficient to have a high probability of exhibiting desired biological activity on the target. Proposed compounds also have to be amenable to chemical synthesis. From the early days of de novo design, it is well known that a major objective for de novo design is the ease of synthesis of the virtually constructed molecules. Nevertheless, for a long time, this issue has been insufficiently addressed. In fact, chemical feasibility of candidate compounds still remains a problem that is far from being completely solved. A common pattern of the de novo design programs that explicitly consider synthetic tractability is to assemble molecular building blocks by rule-based virtual reaction schemes. For example, suitable building blocks can be obtained by virtual *retro*-synthesis of drug molecules. The same set of reactions is then employed for the assembly of new candidate compounds. It is reasonable to assume that such designed compounds will have some degree of "drug-likeness" and contain only few undesirable (e.g., reactive, toxic) structural elements [29]. Virtual structure assembly may be guided by simulated organic synthesis steps so that a synthesis route can be proposed for each generated structure. Alternatively, ligand candidates constructed by de novo design programs can be automatically analyzed by additional software to propose generalized synthetic routes and pick potential reagents from databases of available compounds.

2. Concepts of Virtual Molecule Assembly and Construction of Synthetically Feasible Molecules

A critical step to keep control of compound accessibility by real chemical synthesis is the way the software assembles a compound. While atom-based approaches offer a wider space of potential solutions, they are more prone to delivering chemically instable or implausible suggestions. With respect to chemical feasibility, fragment-based approaches are therefore considered to be the more suitable choice. It may be argued that this is the major reason why – to the best of our knowledge – the last purely atom-based de novo design program (RASSE [30]) was published over a decade ago. Nevertheless, the use of molecular fragments instead of atoms as building blocks alone does not guarantee the construction of virtual compounds that are actually amenable to synthesis. Careful selection of fragments and a set of rules that are tailored to avoid the formation of instable connections

are required to enhance chemical accessibility of the proposed structures. The ultimate goal of de novo design, of course, remains to suggest molecules with a desired activity on a pharmacological target. As a consequence, fragment libraries and connection rules should be compiled under the precondition of keeping the space of constructible molecules as unconstraint and universal as possible. The following sections introduce several approaches to in silico molecule construction, with a particular focus on their ability to yield synthetically feasible candidate structures.

2.1. Alignment-Based Methods

Several de novo programs make use of known binding poses of different ligands bound to the same active site of a target protein (or of closely related proteins). These ligands can be aligned by overlaying the protein structures according to backbone atom coordinates. The resulting multiple alignment of all co-crystallized ligands can then be used to suggest new candidate compounds by combining fragments from different ligands in an automated fashion. The software BREED [30], for example, detects bonds from two different ligands that are in close proximity in their spatial alignment. Geometric constraints regarding the relative orientation of bonds and coordinates of end point atoms are employed to scan for potential break points. Pairs of bonds fulfilling the constraints are cleaved in both compounds A and B, resulting in four fragments A1, A2, B1, and B2. These fragments are recombined to form two new "chimeric" compounds by connecting fragment A1 with B2, and fragment B1 with A2, respectively (Fig. 1).

A related strategy named *fragment shuffling* has recently been developed by researchers at Bayer [17]. While its basic idea corresponds to the one of BREED, it features two extensions: Fragment shuffling (1) is able to combine fragments from more than two ligands within one step, and (2) performs an additional check for steric clashes prior to fragment fusion. While BREED only puts little focus on the accessibility of the created compounds *via* real chemical synthesis (all non-ring bonds of the same bond order lying within the geometrical restraints are considered cleavable), the fragment shuffling strategy contains a set of chemically motivated rules for the selection of cleavable bonds.

An approach picking up the basic idea of BREED to connect fragments by detection of overlapping bonds is implemented in the MED-Hybridise software [15]. Instead of aligning complete ligands, MED-Hybridise first populates the target binding site with predefined fragments termed *MED-Portions* (another strategy that populates the binding site with fragments is MCSS [31]). A MED-Portion is a molecular fragment together with a small patch of a protein surface obtained by a precalculated analysis of known protein–ligand complexes from the Protein Data Bank

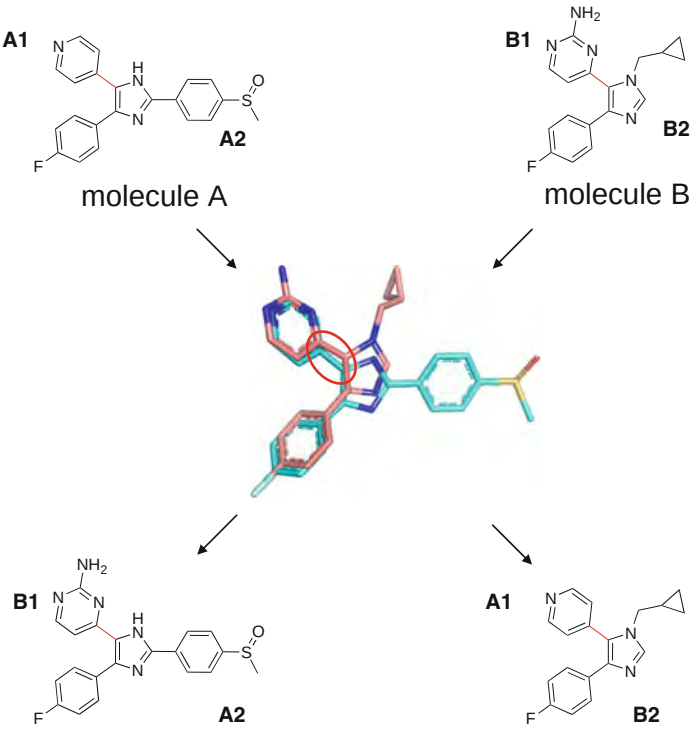


Fig. 1. Cleavable bonds (*red*) are identified within the alignment (*red circle*, *center*). Molecules A and B are divided at these bonds in two fragments each (i.e., A1, A2, B1 and B2, *top*). These fragments are recombined to form two new hybrid compounds (*bottom*).

(PDB) [32]: Each ligand is divided into (possibly overlapping) fragments by searching for substructure matches with available small compounds from the PubChem database [33]. In case of a match, the fragment is stored as a MED-Portion together with a close part of the presumably corresponding protein surface. Protein surface patches are stored as a topology preserving graph of triplets of potential interaction points. These surface patches are used during a productive design run to position MED-Portion fragments within the binding site by searching for matches between their graph representation and the graph constructed of the pocket surface (Fig. 2). Fragments are then connected following the BREED strategy by scanning for overlapping bonds. Unlike BREED and fragment shuffling, this approach assumes only a single 3D model of the binding site and does not depend on the availability of a series of protein–ligand complexes of the same or closely related targets. Since MED-Hybridise employs a fragment definition based on available compounds, the chances of obtaining chemically feasible compounds are increased.

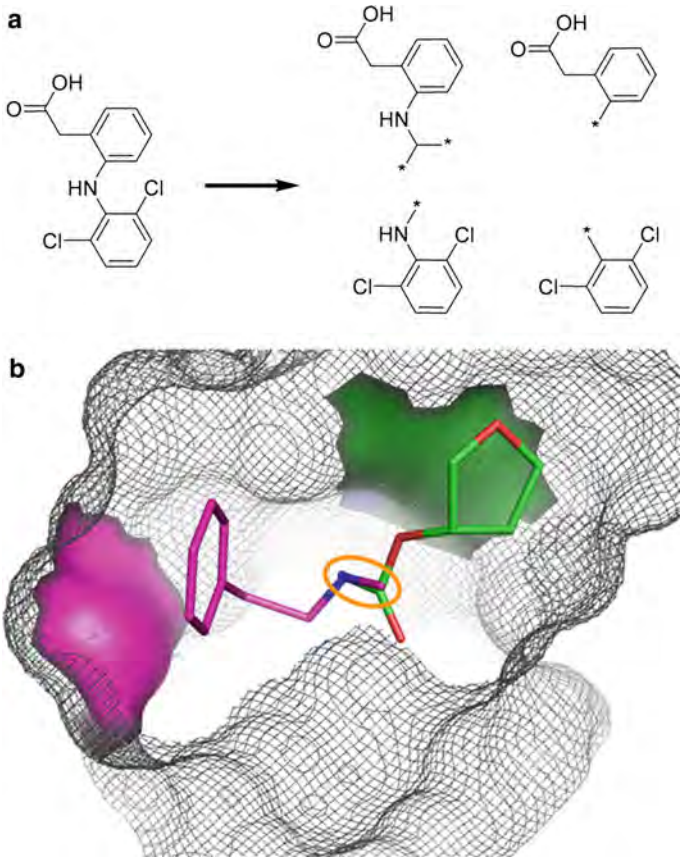


Fig. 2. (**a**) A possible decomposition of Diclofenac (*left*) into substructures defined by commercially available building blocks (*right*; * = attachment site to remaining part of original ligand). (**b**) Surface patches of MED-Portions (*green* and *magenta*; surface patches and corresponding fragments are depicted in the same color) are positioned to match surface properties of the binding pocket (surface shown as mesh). Overlapping bonds are considered for fragment fusion (*orange circle*).

2.2. Methods Based on Molecular Force Fields and Docking

Placing molecular fragments inside a binding cavity in order to connect them in successive steps is a widely used design approach. The software CONCERTS [34] is an early example of using molecular dynamics simulations for fragment placing. Here, fragments are moved according to a molecular force field to obtain low-energetic orientations with respect to interactions with the binding site but without witnessing each other. Bonds can be formed between fragments that are in close proximity. Bonds formed in earlier steps are allowed to break. Such a constant rearrangement of bonds between fragments is supposed to eventually result in compounds exhibiting interaction energies that are favorable to those of the unconnected fragments. Chemical feasibility of designed compounds is supported only by user-defined atom sequences that are disallowed to emerge from formation of new bonds.

Several tools for de novo design make use of docking software in order to initially place fragments into a binding site. In general, two different strategies exist for this approach, i.e., *growing* and *linking* strategies. Growing approaches [35–43] to fragment assembly start with one fragment that already satisfies key interactions with the receptor and add more fragments step-by-step in order to improve ligand affinity, guided by the scoring function of the underlying docking program. The linking strategy [44–50] first places several fragments at distinct parts of the pocket, which are then connected to each other by linker fragments. Again, both approaches typically account for synthetic accessibility only by connection rules defining potential attachment points and disallowed bond formations.

2.3. Fragment Assembly Based on Connection Statistics

In a recent publication, Kutchukian et al. [14] proposed an algorithm termed FOG (Fragment Optimized Growth), which builds up new molecules by an intuitive approach. FOG is based on a set of predefined molecular fragments. In a first step, the algorithm counts how often each fragment is connected to every other fragment in a set of compounds with desired properties (training set). Obtained counts are then converted to connection probabilities. These *transition probabilities* form the basis of the growth strategy implemented as a Markov chain of first order: Following the idea of a Markov chain [51], the process of growing a molecule can be seen as a graph where each fragment is represented by a node. Edges between nodes represent the possibility to connect respective fragments by a molecular bond. Each edge is associated with the corresponding transition probability. To grow a molecule, the algorithm starts with a randomly selected or given fragment (a node) and decides which fragment to add next (which edge to take) based on precalculated transition probabilities. Since each node represents exactly one fragment, transition probabilities only depend on the fragment to be extended in the next step (first order property of the Markov chain). The process stops when all potential growth sites are saturated, a user defined maximum number of fragments is reached, or a maximum molecular mass is exceeded. The Markov chain is supposed to generate molecules that reproduce connection statistics of the training set, therefore exhibiting increased probability to show desired molecular properties.

Two limitations of FOG are that only single bonds are formed between fragments and ring closures are completely forbidden. This certainly limits the search space and restricts the diversity of chemotypes that can be generated. To increase the synthetic feasibility of designed compounds, a set of disallowed 3mers (a 3mer is a sequence of three fragments connected in a row) is defined describing known unwanted or instable substructures and 3mers unseen in the training set. In the presented study based on a small

set of fragments, the authors showed that proposed structures receive SYLVIA scores for synthetic accessibility (vide infra) similar to the scores of training set molecules. An additional survey of organic chemists demonstrated that none of the FOG molecules was deemed unsynthesizable or instable.

2.4. Retrosynthetic Rules for Disassembly and Reconstruction

Virtual retrosynthesis rules find application in some de novo software tools to guide the mining of fragment building blocks as well as reassembly of novel structures in a combinatorial fashion. The most prominent representative of *retro*-synthetic rules is the *Retrosynthetic Combinatorial Analysis Procedure* (RECAP) [52]. RECAP derives 11 cleavable bond types from common chemical reactions and defines them by their molecular environment (Fig. 3). The program TOPAS [29] and its direct successor FLUX [6, 7] employ RECAP to first disconnect a collection of bioactive molecules into nonoverlapping building blocks (fragments). For each cleaved bond, the type of applied rule is stored at the so-generated attachment sites of the fragments. This information is used during fragments assembly, so that only attachment points of the same type are allowed to be reconnected. This approach is supposed to breed druglike molecules that are more stable and synthesizable since the chemical environment around a newly formed bond is the same (at least to some extent) as it is in known drugs and lead compounds. Although this is a more ambitious approach to account for chemical stability and feasibility than the methods mentioned so far, it still bears some critical caveats:

1. The rules only cover a tiny fraction of chemical reactions. For example, no ring formation rule is defined by the original RECAP, thus limiting novel structures to be recombinations of existing ring structures.

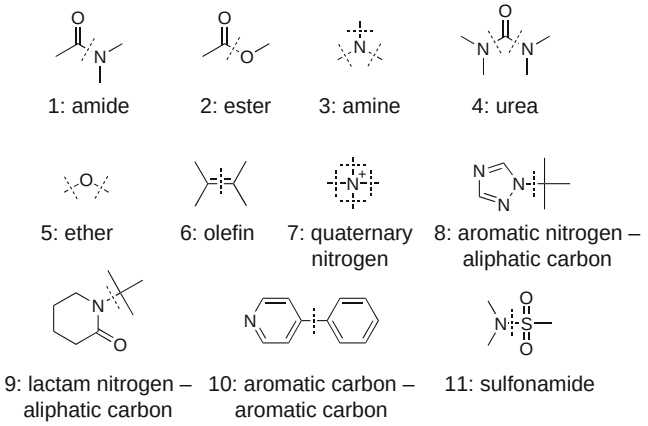


Fig. 3. Eleven RECAP (Retrosynthetic Combinatorial Analysis Procedure) bond cleavage rules [52].

2. RECAP rules are crude abstractions of actual chemical reactions summarizing information of a whole class of reactions and do not represent actual reactions themselves.
3. The approach of dis- and re-assembly does not deliver direct suggestions for synthesis routes. It is not guaranteed that a bond cleaved by a RECAP rule has been formed by a respective reaction that follows the rule's general idea in the sense of educts and product. Building blocks emerging from *retro*-synthetic cleavage can not necessarily be expected to have a direct or even a closely related physically available analog.

2.5. Reaction-Based Approaches

All approaches mentioned so far try to account for synthetic accessibility on a comparably low level. Additional software that especially focuses on the ease of synthesis can introduce an additional scoring function, like SYLVIA [53], which computes a weighted sum from five subscores: (1) a graph complexity score that comprises size, degree of branching, symmetry, bond- and atom-type composition of the molecular graph, (2) a ring complexity score penalizing fused and bridged ring structures, (3) count of stereo centers, (4) a score that estimates how well the structure is covered by available starting materials, and (5) a score that evaluates how well a structure can be disassembled by retrosynthetic analysis to obtain potential synthesis routes. Individual weights for these five scores were calculated by a linear regression analysis to maximize the correlation of final SYLVIA scores with those derived by a survey of medicinal chemists.

Although this approach can explicitly incorporate synthetic feasibility as a design objective into the overall score, it does not supply the user with explicit synthesis routes for suggested compounds. An alternative is to apply additional software for synthesis planning like CAESA [43] or Route Designer [54] after a de novo design run.

The most sophisticated way to explicitly consider ease of synthesis and pursuable synthesis routes directly during construction is to use known chemical reactions as connection rules and readily available molecular building blocks as fragments. An example of software following this idea is the program SYNOPSIS [55]. SYNOPSIS works on a subset of the ACD database [56] and a collection of 70 selected organic reactions. Additional rules are implemented to further incorporate chemical knowledge and guide the acceptance of a specific reaction. These rules consider the molecular neighborhood of a reactive functional group, and the presence of other functional groups that might hinder the reaction or define reactivity preferences when a reactive functional group is present more than once. Beyond that, the authors showed that quantum mechanic calculations can be

successfully applied to predict substituent effects on the reactivity of aromatic halogens. Although it was concluded that these calculations are too time consuming for practical use, the example demonstrates that quantum mechanic calculations have the potential to more generally consider substituent effects and predict reactivity. Certainly, future hardware improvements will enable a wider use of such computationally expensive methods in de novo design.

Simulation of chemical reactions on the computer requires a formalism that is capable to encode substructures of educts directly participating in the reaction (*reaction center*) as well as bond rearrangements caused by the reaction to form the product. Well-established data formats used in commercial software packages to describe reactions are the SMIRKS language [57] and rxn files [56]. Recently Patel et al. proposed a method to automatically extract information from chemical reactions stored in reaction databases as rxn files and convert them to so called *reaction vectors* [58]. A reaction vector is a generalized version of a reaction, limited to participating reaction centers and a confined environment of surrounding atoms. Generalization makes reaction vectors applicable to simulate reactions on a broad range of starting materials for de novo compound design. While extracting information in an automated fashion has the virtue to easily create huge data pools, it generally bears the caveat of having little control over the quality of the collected data. Manual post processing is required prior to application, shortening some of the advantages of automated data mining.

We are currently working on the new de novo design software DOGS (Design of Genuine Structures). Similar to SYNOPSIS, DOGS is based on a manually selected set of reactions collected from literature. Selection criteria comprised ease of application, high yields, and generation of substructures commonly found in druglike molecules. Special focus was put on reactions forming new ring structures as central scaffolds. Reactions are encoded using our own formal reaction language *Reaction-MQL* [59]. Available starting materials from several vendor catalogues are employed to construct compounds that feature high graph similarity to a known reference ligand. Graph similarity is measured by means of the ISOAK graph kernel method [60] either directly on the molecular graph or on a reduced graph in order to account for a higher level of abstraction from the atomic structure. For each compound designed by DOGS, at least one potential synthesis route is proposed. Figures 4 and 5 present examples of compounds constructed by DOGS together with suggested synthesis routes for two different reference ligands.

A feature of DOGS rarely found among de novo software tools is the complete avoidance of any stochastic optimization (other examples of deterministic optimization are FlexNovo [8]

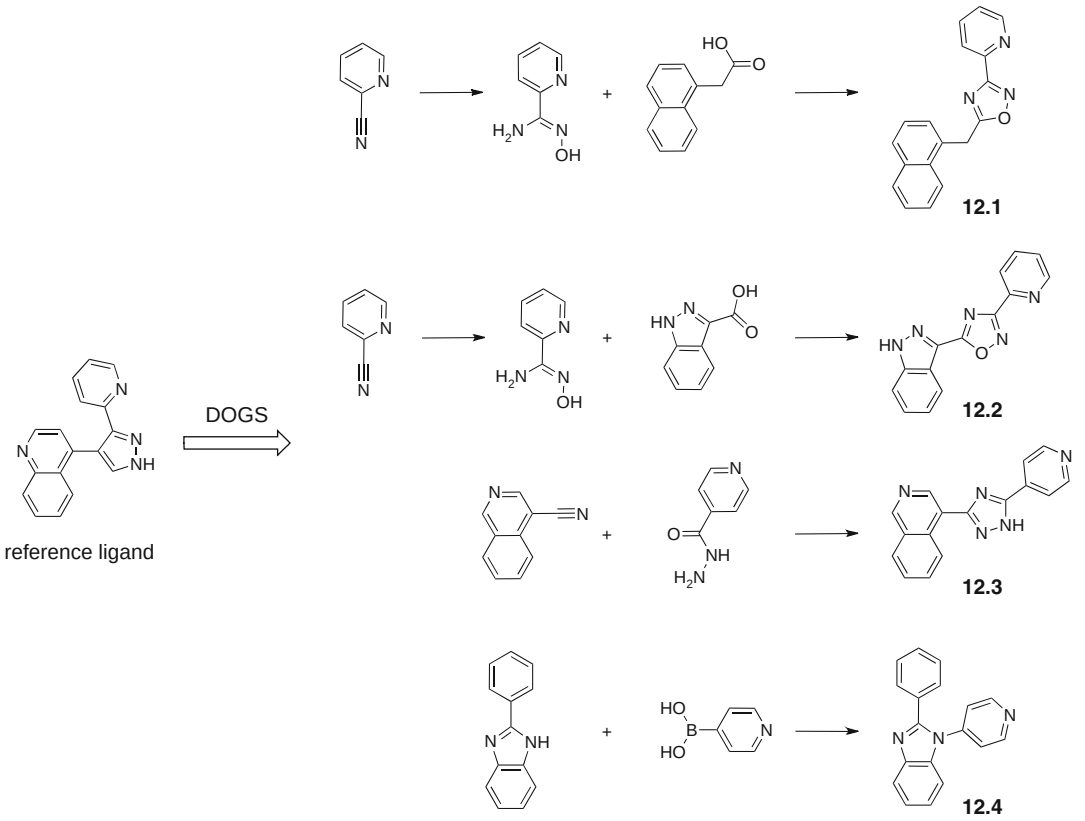


Fig. 4. Compounds **12.1–12.4** and respective hypothetical synthesis routes were suggested by DOGS, given a potent TGFβ (transforming growth factor β) inhibitor as reference ligand (*left*).

and fragment shuffling [17]). This is achieved by implementing a greedy depth-first search heuristic in combination with full enumeration of parts of the search space. At this point DOGS benefits from an effect of its "real" chemistry approach: Search spaces are restricted to a more manageable size in a practice-oriented way. This cuts down the sizes of subspaces, which need to be exhaustively scanned by DOGS.

Ligand assembly is a step where control about accessibility *via* chemical synthesis can be easily addressed. Nevertheless, only a small fraction of available de novo software tools put effort in this design task in a nontrivial way. Suggesting synthesizable molecules is not only important for successful application of design software in drug discovery projects but also for validation of the program. As long as designed molecules remain virtual and not practically tested in the laboratory due to the lack of a feasible synthetic route, validation remains theoretical – a problem de novo design has been afflicted with since its early days.

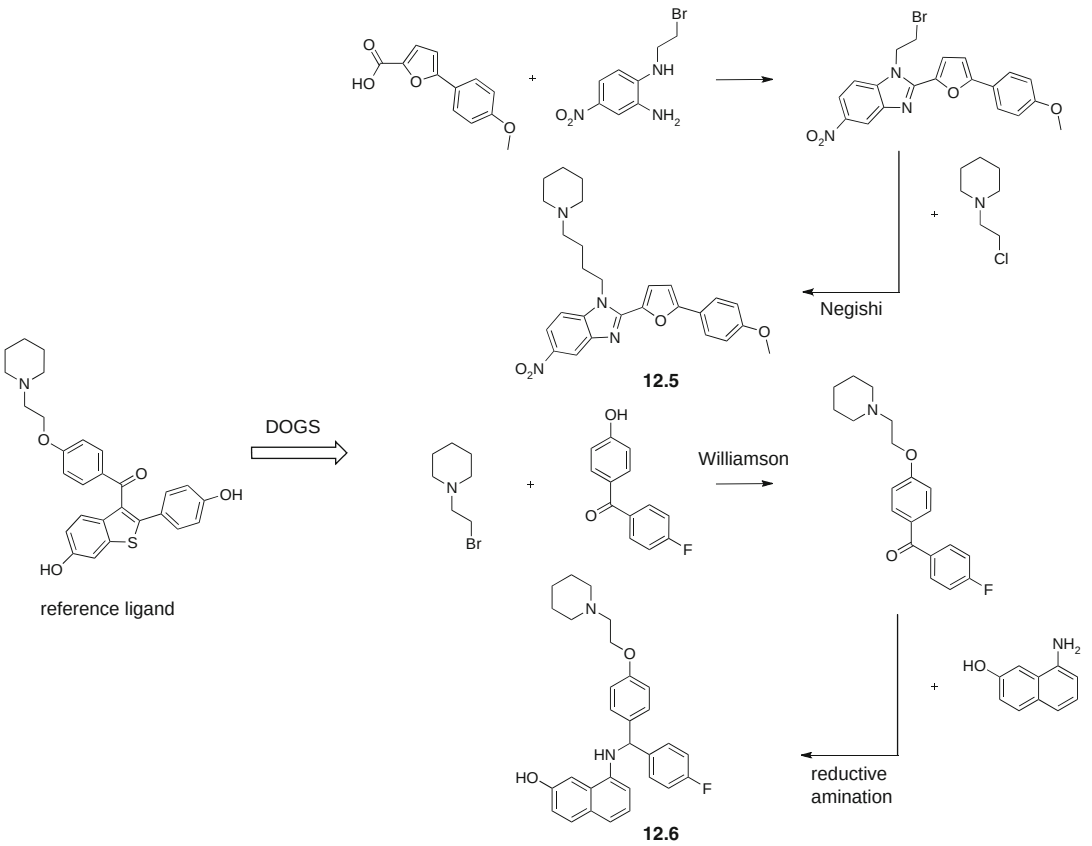


Fig. 5. Compounds **12.5** and **12.6** together with hypothetical synthesis routes were suggested by DOGS, given the selective estrogen receptor modulator Raloxifen as reference ligand (*left*). Arrows are annotated with reaction names in cases where general names exist.

3. Fusing Virtual Screening with De Novo Design

De novo design usually has to deal with large search spaces due to combinatorial explosion [2, 14, 21]. The number of molecules in these search spaces easily exceeds the limits of today's computational power by orders of magnitude [61, 62], which only too often prohibits exhaustive enumeration of all possible solutions. Many programs deal with this problem by applying stochastic search strategies like Evolutionary Algorithms, Metropolis Search or Simulated Annealing [20]. While such stochastic processes have shown to bring up results of high practical relevance even in very large search spaces, they cannot guarantee finding the best solution. In fact, given the high complexity of molecular design, there is a high probability that two runs of the same stochastic optimization process will end up with different results. On the other side, this ambiguity may also be exploited to generate structurally

diverse sets of candidate compounds and avoid over-optimization and artifacts due to imperfect scoring functions [20].

The (theoretical) need to explicitly enumerate all possible combinations of fragments to guarantee for finding the overall best solutions is a consequence of the scoring schemes employed by de novo software tools. Strictly speaking, scores should be computed for complete molecules and only in a limited number of cases may be calculated as the sum of fragment scores. This is due to the fact that binding energies of ligand–receptor interactions cannot be expected to be additive *per se* [63], although approximately additive substituent effects have been observed in compound design studies [64]. This implies that with a hypothetical perfect scoring function at hand, it would be impossible to avoid the need to enumerate all possible combinations of fragments to find the best solutions for a given de novo design problem. Since we cannot hope to find a perfect scoring function, we must rely on imperfect models for scoring candidate compounds. Every model has the inherent quality to miss some aspects of reality in order to stay manageable. Modeling a scoring function explicitly featuring additivity has the striking advantage that it avoids the need to score the large amount of all possible fragment combinations. Instead, each entry of the much smaller fragment library can be scored and an optimal combination of fragments can be computed that forms the overall best solution. One has to keep in mind that *best* here means *best in the sense of the model*, ignoring its simplifications.

To illustrate the advantage of additive scoring schemes, consider two libraries of complementary fragments for combinatorial chemistry each with 1,000 entries. Complete enumeration of all possible products would result in 1,000,000 ($1{,}000^2$) molecules, which all have to be scored. An additive scoring scheme would only have to score each fragment once, resulting in 2,000 (i.e., 1,000 + 1,000) scores. The advantage of additive scoring increases as the number of fragments rises.

Some de novo software tools make use of scoring functions, which are additive in nature. For example, the fragment shuffling method [17] (vide supra) uses atom scores from the FlexX docking software [65] to calculate additive fragment scores. Fragment scores are then employed to guide a greedy tree search identifying the most promising combinations of fragments for hybridization while avoiding construction of unfavorable ones. Another example is a program proposed by Nikitin et al. [4] (vide infra). Here, each side chain of a preselected scaffold can be optimized independently from all other side chains by utilizing an additive scoring scheme, turning this originally multiplicative problem into an (simplifying and approximately) additive one.

Additive scoring schemes also build the basis of algorithms bridging conventional virtual screening and fragment based de

novo design. Two approaches following this idea are the topomer search proposed by Cramer et al. [66] and FTrees-FS (Feature Trees Fragment Space Search) [67]. The first step of a topomer search is the fragmentation of a reference ligand according to an existing synthesis route. Steric field properties of each reference fragment are computed and compared to precalculated steric fields of a special topomer library (one per generated fragment). A topomer library comprises readily available starting materials, which comply with the structural requirements of the underlying synthesis route, i.e., presence of required reactive substructures. From each topomer library, fragments are chosen that have the strongest shape similarity to the respective reference fragment and forwarded to a real combinatorial synthesis. Since the employed steric field similarity measurement is additive, the overall shape similarity of connected molecular fragments (a designed candidate compound) to the reference ligand computes as the sum of the topomer similarities of the fragments. Therefore, it is not necessary to explicitly enumerate all fragment combinations prior to scoring. This reduces the number of necessary comparisons dramatically while still being guaranteed to find the overall best solutions.

A similar approach is taken by FTrees-FS: Like in topomer search, FTrees-FS encodes the large space of possible products implicitly by means of a compact fragment collection instead of explicitly enumerating them exhaustively. The main difference to the topomer search is that the reference ligand is not cleaved into fragments to search for suitable bioisosteric replacements for each fragment separately. Instead, FTrees-FS directly constructs complete products and compares them to the reference molecule as a whole. Construction is guided by a dynamic programming algorithm together with an additive similarity scoring function based on the FeatureTrees representation [68]. The algorithm is guaranteed to construct those virtual products exceeding a given similarity threshold without the need to construct all possible product structures of the combinatorial space. FTrees-FS works on the 2D molecular structure instead of 3D conformers, making a screen of the search space considerably faster than a topomer search (minutes for a theoretical size of 10^{12} instead of hours for 10^{9} molecules). While the first implementation of FTrees-FS [67] worked on RECAP with the aforementioned drawbacks regarding chemical feasibility of products, recent developments take existing combinatorial chemistry protocols and available starting materials as the basis of construction [69, 70].

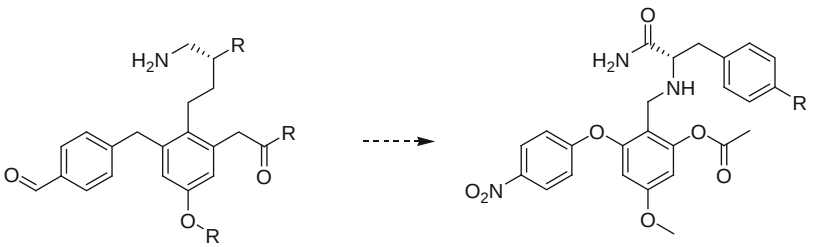
Both methods described combine features of de novo design (storage of product space in its closed form as fragment building blocks and connection rules) and virtual screening (finding the overall best solutions like an exhaustive search). Despite the appeal of exhaustive and deterministic searching, which can be enabled

by additive scoring, it should be kept in mind that such an approach is particularly prone to errors in the scoring function, which requires most careful design.

4. Recent Applications of De Novo Design Software

Compared to virtual screening, the number of reported successful de novo design campaigns is significantly smaller. The reason clearly can be addressed to the higher complexity and costs of de novo studies: Besides computational efforts and costs for ordering and testing of compounds, chemical syntheses have to be carried out causing a considerable amount of work, financial investment, and presuming additional expert knowledge. Hence, de novo design cannot be considered a standard tool in computer aided drug design yet. A more psychological aspect is that whenever a software suggests compounds that resemble already known drugs or leads, there is little enthusiasm to synthesize such molecules (although it may be considered a successful design run). As soon as the designs are more unexpected, there is an understandable reluctance to synthesize such a molecule, because it may well be inactive, and custom synthesis can be time-consuming and requires significant man power. Still, reports are increasing which present successful applications of de novo design software in drug discovery. Here, we will review selected recent examples.

The software SPROUT [41–43] has recently been adopted in two successful design studies. On the basis of an X-ray crystal structure of *Escherichia coli* RNA polymerase, Agarwal et al. reported the design of a template structure with the help of SPROUT (Fig. 6, left panel) [71]. Successive modifications and straightforward synthesis lead to the two close analogs **12.7** and **12.8** with moderate activity on the target protein (IC_{50} values: 70



template designed by SPROUT

12.7: R = OH
12.8: R = H

Fig. 6. Molecules **12.7** and **12.8** were synthesized as close analogs of a template structure suggested by SPROUT [41–43] (*left*) for the binding site of *E. coli* RNA polymerase.

and 62 μM, Fig. 6, right panel). The authors stress that ligand design for this target is especially challenging because of its fairly open and solvent exposed cavity. In addition, it has been reported that HTS methods have failed on the same target, demonstrating that de novo design can be a valuable tool complementary to standard approaches.

In another study, SPROUT was applied to suggest new inhibitors of ligase VanA from *Enterococcus faecium* [72]. A known bioisosteric analog for the high-energy intermediate of the reaction catalyzed by the enzyme was chosen to form the basis of novel candidate compounds (Fig. 7, center). The two most promising structures were chosen for synthesis, with a small modification comprising the replacement of a tetrahydrofuryl ring by morpholine (Fig. 7, right, dashed circle). This small structural change keeps the potential ability to form a hydrogen bond with the receptor suggested by the software but simplifies chemical synthesis. Although IC_{50} values are in the high micromolar range (**12.9**: 224 μM, **12.10**: 290 μM), these compounds may represent potential starting points for further optimization. The ligands are of particular interest as they exhibit inhibitory activity on another ligase of *Escherichia coli* as well, which makes them precursors of potential dual inhibitors of ligases from both Gram-positive and Gram-negative bacteria.

In 2005 Nikitin et al. proposed an implementation of a software tool for receptor-based de novo design together with its successful application on human immunodeficiency virus (HIV) integrase [4]. Out of 22 compounds chosen for synthesis from the result list returned by the software, 20 were synthesized without major difficulties. Among those were compounds featuring both known and unknown scaffolds for the target. The authors found several of the 20 structures to have weak activity values. One previously unknown compound of a known active scaffold yielded

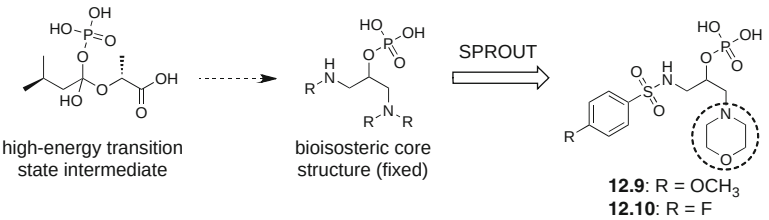
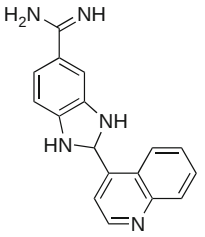


Fig. 7. Molecules **12.9** and **12.10** were synthesized after a minor modification of the structures designed by SPROUT [41–43] to inhibit ligase VanA from *Enterococcus faecium*. Modifications comprised the exchange of a tetrahydrofuryl ring against a morpholine moiety (*dashed circle*, *right*). SPROUT design was based on a fixed core structure (*center*) known to be a bioisoster for the high-energy transition intermediate of the reaction catalyzed by the enzyme (*left*).



12.11

Fig. 8. An inhibitor of HIV integrase synthesized as suggested by an algorithm proposed by Nikitin et al. [4]. The benzimidazole scaffold introduces a new potential entry point for further investigations on new inhibitors of HIV integrase.

an IC_{50} of 2.2 μM. In addition, compound **12.11** (Fig. 8) was found to inhibit HIV integrase with an IC_{50} of 15 μM selectively compared to some other strand-breaking enzymes. The benzimidazole moiety of compound **12.11** introduces a scaffold previously unknown for HIV integrase inhibitors. The compound leaves room for further optimization, which is a desired property of de novo designed ligands.

In a study, Park et al. [73] demonstrate how knowledge about active scaffolds can be directly incorporated into de novo design. The authors used the program LigBuilder [74] for automated design of inhibitors for human Cdc25 phosphatase. Design runs were based on two promising scaffolds that had been identified by virtual screening. An additional term explicitly considering for solvation effects during docking was introduced to the scoring function of LigBuilder. Of 10,000 structures generated in silico, the top scoring 500 for each scaffold were selected and checked for commercial availability: 107 (scaffold 1) and 82 (scaffold 2) compounds were purchased and tested against subtypes A and B of Cdc25 phosphatase. For scaffold 1, 19 ligands exhibited an IC_{50} < 31 μM for both subtypes with the overall best inhibitor **12.12** yielding $IC_{50} = 5.1$ μM for subtype A and 1.2 μM for subtype B (Fig. 9, top panel). Among the structures derived from scaffold 2, 13 inhibitors had an IC_{50} below 20 μM for both subtypes (best inhibitor **12.13**: 4.7 μM on subtype A and 2.3 μM on subtype B; *cf.* Fig. 9, bottom panel). This study is somewhat atypical for de novo ligand design since (1) a considerable part of the molecules was kept fixed during the design, and (2) the designed compounds were purchased instead of being synthesized. Although this foils the basic scope of de novo design to construct structures unseen before, it demonstrates that the algorithm is able to propose chemically stable and active molecules once a feasible scaffold is identified – either by the software itself or by intervention of the user. This assertion is well supported for the presented study by the comparably high number of molecules tested and can be extended to cases where the software is applied in more realistic

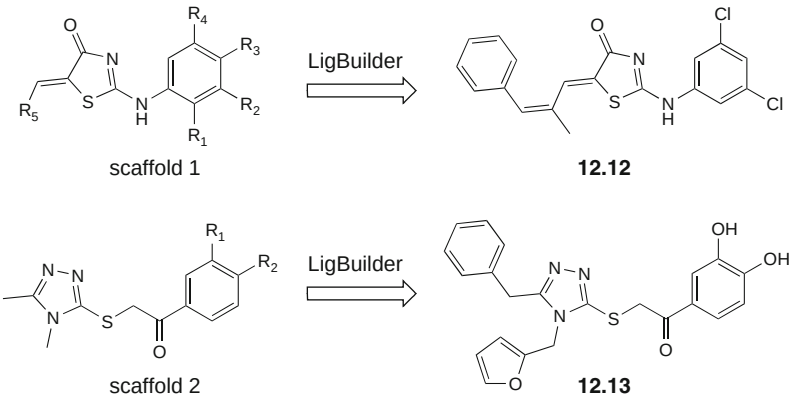


Fig. 9 LigBuilder [74] designed active inhibitors **12.12** and **12.13** of the human Cdc25 phosphatase, each representing the most potent structure of two design series based on one of the scaffolds 1 and 2 (*left*).

de novo design scenarios. Usually, the number of tested molecules from de novo design software is considerably smaller (1–10) due to expensive syntheses and as a consequence of insufficient statistics statements about expected general quality of the output remain more speculative.

All applications presented so far have been based on 3D structural models of protein targets. Ligand-based de novo design for an RNA target has been reported by Schüller et al. [75]. In this study, the software FLUX [6, 7] was employed to design novel ligands of the *trans*-activation response element (TAR) of the HIV-1 mRNA, which is a potential drug target in the treatment of AIDS. Although 3D models of this RNA exist, TAR is known to be exceptionally flexible and therefore a challenging target for drug design, especially for methods using information from the binding site. FLUX, as a purely ligand-based approach, does not exploit information about the 3D structure of the target and hence it is suited for such a design task. Acetylpromazine served as reference ligand to guide the construction of new candidate compounds by FLUX. The aim was to design structures with minimum distance to Acetylpromazine with respect to the relative distribution of pharmacophoric features encoded by the CATS descriptor [76]. The top scoring compound **12.14** was slightly modified after inspection by a chemist to obtain compound **12.15** (Fig. 10). Modifications comprised the exchange of a tetrafluoroethoxy moiety with an ethoxy group to facilitate synthesis and increase drug-likeness. Synthesis of compound **12.15** revealed that it has the desired property to inhibit the interaction between TAR and the binding protein, albeit at relatively high concentrations. This work is of particular interest as it represents the first successful application of automated de novo design for an RNA target.

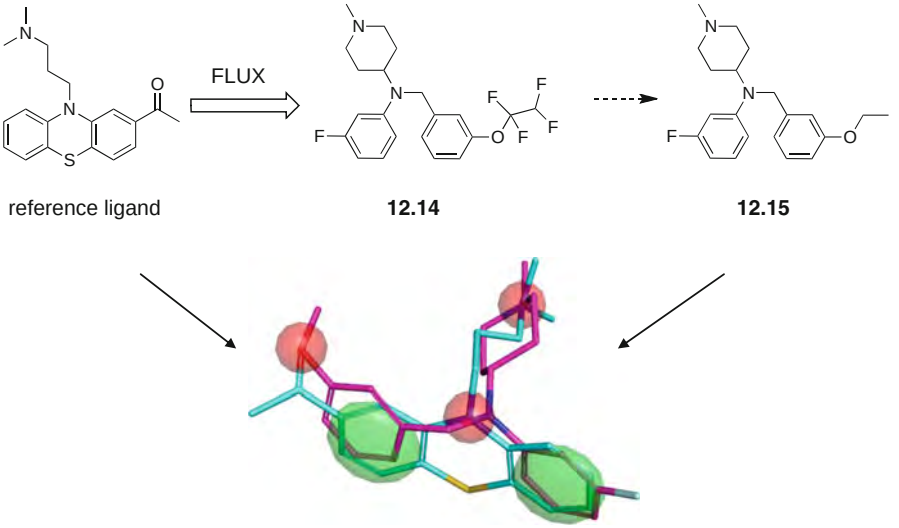


Fig. 10. Compound **12.15** was synthesized and showed inhibition of the interaction between TAR element of HIV-1 mRNA and the tat protein. **12.15** is a close analog of structure **12.14** designed by a FLUX [6, 7] de novo design run based on the known reference ligand Acetylpromazine (*top panel*, *left*). A flexible alignment of the reference compound and **12.15** (*bottom panel*) shows good agreement in steric and pharmacophor properties (*green*: aromatic centers, *red*: hydrogen bond acceptor).

5. Conclusions and Outlook

Although it has recently been stated that no major breakthroughs have been made in the field of automated de novo drug design in the past 2 years [77], we are currently witnessing an increased interest in this field. Fragment-based approaches in combination with assembly strategies based on knowledge about chemical reactions restrict the vast room of compounds constructible on a computer to reasonable subspaces of synthesizable structures of interest from the view of medicinal chemistry. This does not only help deal with huge potential search spaces but also increases the chance to remedy a problem de novo design is confronted with from its early days: the lack of practical validation. Synthesis and testing of compounds designed in silico is the only way to truly assess the value of de novo design algorithms for drug discovery. Incorporation of knowledge about chemical synthesis is – at least from a technical point of view – no longer a problem.

Scoring functions with (approximately) additive fragment scores [4, 17, 66, 67] represent a trend of the last few years that complements stochastic search strategies used in earlier approaches. While additive scoring characteristics are a desired property to deal with huge numbers of potential compounds, the most important property of a good scoring function is its

accuracy in distinguishing active compounds from inactive ones. Here, the main shortcomings of de novo design are obvious: Receptor-based scoring of receptor–ligand complexes does not predict measured activity data with sufficient accuracy. A main reason is the still inadequate consideration of entropic effects of receptor–ligand interactions caused by flexibility and solvent effects. While flexibility of the ligand is explicitly considered by almost all docking procedures, its effects on binding energies are only coarsely estimated in most cases. Many docking algorithms keep the receptor fixed and completely disregard entropic effects of the binding pocket and induced fit behavior. Two approaches that incorporate conformational flexibility of the receptor into de novo design to account for induced fit phenomena within the software SkelGen have been published recently. Todorov et al. [78] took a series of different static receptor structures as the basis for a design run with SkelGen instead of only a single one to account for potential flexibility of the binding site. Another approach has been followed by Alberts et al. who allowed for side chain movements of selected protein residues by either precalculated combinations of rotamers or sampling of torsion angles [79]. However, no scoring function used for de novo design so far explicitly computes entropic effects of the solvent caused by ligand binding. Improvement in the accuracy of scoring functions is a field where we see great potential for future breakthroughs in de novo design.

During the last few years, some approaches were published that are closely related to de novo design while breaking with the basic idea of this field to start completely from scratch. These algorithms can be seen as variations of "pure" de novo design. They start with knowledge about substructures of available inhibitors and only add what is needed for completion as ligand candidates. This can be achieved by bridging cyclic fragments already placed within the binding site (either by docking or knowledge from 3D structure elucidation), by special bridge fragments as implemented in the CONFIRM software [50], or by replacing core structures of known ligands with other suitable scaffolds. The latter approach is taken by the software Recore [80], which uses knowledge about the binding mode of a known ligand to perform scaffold hopping by searching for different scaffolds with the potential to keep the side chains of the old ligand in position and additionally interact with the receptor itself. Grabowski et al. demonstrated that this concept is able to find simplistic bioisosteric replacements for more complex natural product cores [81]. The program COLIBREE [82] takes exactly the complementary approach and takes a user-defined scaffold as input to build a focused library of potential ligands. The given scaffold is held fixed during design while only side chains at designated positions of the scaffold are optimized to maximize similarity to given

reference ligands, which do not necessarily have to possess the scaffold themselves. The possibility to define a fixed core template has been implemented as an additional option in de novo software tools like SPROUT and LigBuilder, which also have the ability to completely start from scratch (Figs. 7 and 9).

Summarizing, we see an increased interest in de novo design for rescaffolding and lead structure generation. Several practical validation studies have already demonstrated that rule-based fragment assembly can result in novel synthesizable compounds with druglike properties and a desired activity. Although successful steps have been made toward the emulation of realistic compound synthesis, the available software tools are far from being able to automatically design ready-to-use lead compounds. While this should be considered as the ultimate goal, current design techniques already enrich drug discovery projects by suggesting new chemotypes. Advances in chemical synthesis such as microreactors and lab-on-a-chip approaches will undoubtedly stimulate computational chemistry so that innovative de novo design software will be developed. We are convinced that there is much more to expect from de novo design in the future.

Acknowledgments

The authors thank Herbert Köppen and Karl-Heinz Baringhaus for stimulating discussion. M.H. is grateful to Merz Pharmaceuticals for a scholarship. This research was supported by the Beilstein Institut zur Förderung der Chemischen Wissenschaften and the DFG (SFB579, project A11.2).

References

1. Danziger, D. J. and Dean, P. M. (1989) Automated site-directed drug design: a general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc R Soc Lond B Biol Sci* **236**, 101–13.
2. Schneider, G. and Fechner, U. (2005) Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* **4**, 649–63.
3. Mauser, H. and Guba, W. (2008) Recent developments in de novo design and scaffold hopping. *Curr Opin Drug Discov Devel* **11**, 365–74.
4. Nikitin, S., Zaitseva, N., Demina, O., Solovieva, V., Mazin, E., Mikhalev, S., Smolov, M., Rubinov, A., Vlasov, P., Lepikhin, D., Khachko, D., Fokin, V., Queen, C., and Zosimov, V. (2005) A very large diversity space of synthetically accessible compounds for use with drug design programs. *J Comput Aided Mol Des* **19**, 47–63.
5. Douguet, D., Munier-Lehmann H., Labesse G., and Pochet S. (2005) LEA3D: a computer-aided ligand design for structure-based drug design. *J Med Chem* **48**, 2457–68.
6. Fechner, U. and Schneider, G. (2006) Flux (1): a virtual synthesis scheme for fragment-based de novo design. *J Chem Inf Model* **46**, 699–707.
7. Fechner, U., and Schneider, G. (2007) Flux (2): comparison of molecular mutation and

crossover operators for ligand-based de novo design. *J Chem Inf Model* **47**, 656–67.
8. Degen, J. and Rarey, M. (2006) FlexNovo: structure-based searching in large fragment spaces. *ChemMedChem* **1**, 854–68.
9. Feher, M., Gao, Y., Baber, C., Shirley, W. A., and Saunders, J. (2008) The use of ligand-based de novo design for scaffold hopping and sidechain optimization: two case studies. *Bioorg Med Chem* **16**, 422–7.
10. Dey, F., and Caflisch, A. (2008) Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J Chem Inf Model* **48**, 679–90.
11. Proschak, E., Zettl, H., Tanrikulu, Y., Weisel, M., Kriegl, J. M., Rau, O., Schubert-Zsilavecz, M., and Schneider, G. (2009) From molecular shape to potent bioactive agents I: bioisosteric replacement of molecular fragments. *ChemMedChem* **4**, 41–4.
12. Proschak, E., Sander, K., Zettl, H., Tanrikulu, Y., Rau, O., Schneider, P., Schubert-Zsilavecz, M., Stark, H., and Schneider, G. (2009) From molecular shape to potent bioactive agents II: fragment-based de novo design. *ChemMedChem* **4**, 45–8.
13. Hecht, D. and Fogel, G. B. (2009) A novel in silico approach to drug discovery via computational intelligence. *J Chem Inf Model* **49**, 1105–21.
14. Kutchukian, P. S., Lou, D., and Shakhnovich, E. I. (2009) FOG: fragment optimized growth algorithm for the de novo generation of molecules occupying druglike chemical space. *J Chem Inf Model* **49**, 1630–42.
15. Moriaud, F., Doppelt-Azeroual, O., Martin, L., Oguievetskaia, K., Koch, K., Vorotyntsev, A., Adcock, S. A., and Delfaud, F. (2009) Computational fragment-based approach at PDB scale by protein local similarity. *J Chem Inf Model* **49**, 280–94.
16. Nicolaou, C. A., Apostolakis, J., and Pattichis, C. S. (2009) De novo drug design using multiobjective evolutionary graphs. *J Chem Inf Model* **49**, 295–307.
17. Nisius, B., and Rester, U. (2009) Fragment shuffling: an automated workflow for three-dimensional fragment-based ligand design. *J Chem Inf Model* **49**, 1211–22.
18. Durrant, J. D., Amaro, R. E., and McCammon, J. A. (2009) AutoGrow: a novel algorithm for protein inhibitor design. *Chem Biol Drug Des* **73**, 168–78.
19. Schneider, G. and Baringhaus, K.-H. (2008) *Molecular Design*, Wiley-VCH, Weinheim.
20. Schneider, G., Hartenfeller, M., Reutlinger, M., Tanrikulu, Y., Proschak, E., and Schneider, P. (2009) Voyages to the (un)known: adaptive design of bioactive compounds. *Trends Biotechnol* **27**, 18–26.
21. Green, D. V. (2003) Virtual screening of virtual libraries. *Prog Med Chem* **41**, 61–97.
22. Richardson, J. S. and Richardson, D. C. (1989) The de novo design of protein structures. *Trends Biochem Sci* **14**, 304–9.
23. Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernert, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y., McClain, R. D., and Donlan, M. E. (1992) Looking at proteins: representations, folding, packing, and design. *Biophys J* **63**, 1185–209.
24. Lameijer, E. W., Tromp, R. A., Spanjersberg, R. F., Brussee, J., and Ijzerman, A. P. (2007) Designing active template molecules by combining computational de novo design and human chemist's expertise. *J Med Chem* **50**, 1925–32.
25. Gillet, V. J., Khatib, W., Willett, P., Fleming, P. J. and Green, D. V. (2002) Combinatorial library design using a multiobjective genetic algorithm. *J Chem Inf Comput Sci* **42**, 375–85.
26. Gillet V. J. (2008) New directions in library design and analysis. *Curr Opin Chem Biol* **12**, 372–8.
27. Hopkins A. L., Groom C. R., and Alex A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* **15**, 430–1.
28. Bembenek S. D., Tounge B. A., and Reynolds C. H. (2009) Ligand efficiency and fragment-based drug discovery. *Drug Discov Today* **14**, 278–83.
29. Schneider, G., Lee, M., Stahl, M., and Schneider, P. (2000) De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J Comput Aided Mol Des* **14**, 487–94.
30. Pierce, A. C., Rao, G., and Bemis, G. W. (2004) BREED: generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. *J Med Chem* **47**, 2768–75.
31. Miranker, A. and Karplus, M. (1991) Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins* **11**, 29–34.
32. RCSB Protein Data Bank, http://www.rcsb.org/pdb/ (accessed September 28, 2009).
33. The PubChem Project, http://pubchem.ncbi.nlm.nih.gov/ (accessed September 28, 2009).
34. Pearlman, D. A. and Murcko, M. A. (1996) CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. *J Med Chem* **39**, 1651–63.

35. Luo, Z., Wang, R., and Lai, L. (1996) RASSE: a new method for structure-based drug design. *J Chem Inf Comput Sci* **36**, 1187–94.
36. Bohacek, R. S. and McMartin, C. (1994) Multiple highly diverse structures complementary to enzyme binding sites: results of extensive application of a de novo design method incorporating combinatorial growth. *J Am Chem Soc* **116**, 5560–71.
37. Gillett, V. A., Johnson, A. P., Mata, P., and Sike, S. (1990) Automated structure design in 3D. *Tetrahedron Comput Method* **3**, 681–96.
38. Rotstein, S. H. and Murcko, M. A. (1993) GenStar: a method for de novo drug design. *J Comput Aided Mol Des* **7**, 23–43.
39. DeWitte, R. S. and Shakhnovich, E. I. (1996) SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J Am Chem Soc* **118**, 11733–44.
40. Ishchenko, A. V. and Shakhnovich, E. I. (2002) SMall molecule growth 2001 (SMoG2001): an improved knowledge-based scoring function for protein – ligand interactions. *J Med Chem* **45**, 2770–80.
41. Gillet, V. J., Johnson, A. P., Mata P., Sike, S., and Williams P. (1993) SPROUT: a program for structure generation. *J Comput Aided Mol Des* **7**, 127–53.
42. Gillet, V. J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z. and Johnson, A. P. (1994) SPROUT: recent developments in the de novo design of molecules. *J Comput Aided Mol Des* **34**, 207–17.
43. Gillett, V. J., Myatt, G., Zsoldos, Z., and Johnson, A. P. (1995) SPROUT, HIPPO and CAESA: tools for de novo structure generation and estimation of synthetic accessibility. *Perspect Drug Discov Des* **3**, 34–50.
44. Böhm, H.-J. (1992) The computer program LUDI: a new simple method for the de-novo design of enzyme inhibitors. *J Comput Aided Mol Des* **6**, 61–78.
45. Böhm, H.-J. (1992) LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* **6**, 593–606.
46. Böhm, H.-J. (1993) A novel computational tool for automated structure-based drug design. *Journal of Molecular Recognition*, **6**, 131–7.
47. Böhm, H.-J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des* **8**, 243–56.
48. Böhm, H.-J. (1998). Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* **12**, 309–23.
49. Tschinke, V. and Cohen, N. C. (1993) The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypothesis. *J Med Chem* **36**, 3863–70.
50. Thompson, D. C., Denny, R. A., Nilakantan, R., Humblet, C., Joseph-McCarthy, D., and Feyfant, E. (2008) CONFIRM: connecting fragments found in receptor molecules. *J Comput Aided Mol Des* **22**, 761–72.
51. Markov, A.A., (1971) Extension of the limit theorems of probability theory to a sum of variables connected in a chain. In: Howard, R. (Ed.), *Dynamic Probabilistic Systems, vol. 1*, Markov Chains (reprinted in Appendix B), John Wiley and Sons, Hoboken.
52. Lewell, X. Q., Judd, D., Watson, S., and Hann, M. (1998) RECAP – retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* **38**, 511–22.
53. Boda, K., Seidel, T., and Gasteiger, J. (2007) Structure and reaction based evaluation of synthetic accessibility. *J Comput Aided Mol Des* **21**, 311–25.
54. Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S. Y., Johnson, A. P., Major, S., Wade, R. A., and Ando, H. Y. (2009) Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J Chem Inf Model* **49**, 593–602.
55. Vinkers, H. M., de Jonge, M. R., Daeyaert, F. F., Heeres, J., Koymans, L. M., Lenthe, J. H., Lewi, P. J., Timmerman, H., Van Aken, K., and Janssen, P. A. (2003) SYNOPSIS: SYNthesize and OPtimize system in silico. *J Med Chem* **46**, 2765–73.
56. Symyx Technology Inc., 2440 Camino Ramon, Suite 300, San Ramon, CA 94583, USA.
57. Daylight Chemical Information Systems, Inc., 120 Vantis-Suite 550, Aliso Viejo, CA 92656, USA.
58. Patel, H., Bodkin, M. J., Chen, B., and Gillet, V. J. (2009) Knowledge-based approach to de novo design using reaction vectors. *J Chem Inf Model* **49**, 1163–84.
59. Reisen, F. H., Schneider, G., and Proschak, E. (2009) Reaction-MQL: line notation for functional transformation. *J Chem Inf Model* **49**, 6–12.

60. Rupp, M., Proschak, E., and Schneider, G. (2007) Kernel approach to molecular similarity based on iterative graph similarity. *J Chem Inf Model* **47**, 2280–6.
61. Bohacek R. S., McMartin C., and Guida W. C. (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med Res Rev* **16**, 3–50.
62. Mauser, H. and Stahl, M. (2007) Chemical fragment spaces for de novo design. *J Chem Inf Model* **47**, 318–24.
63. Klebe G. and Böhm H. J. (1997) Energetic and entropic factors determining binding affinity in protein-ligand complexes. *J Recept Signal Transduct Res* **17**, 459–73.
64. Jorissen, R. N., Reddy, G. S., Ali, A., Altman, M. D., Chellappan, S., Anjum, S. G., Tidor, B., Schiffer, C. A., Rana, T. M., and Gilson, M. K. (2009) Additivity in the analysis and design of HIV protease inhibitors. *J Med Chem* **52**,737–54.
65. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **261**, 470–89.
66. Cramer, R. D., Poss, M. A., Hermsmeier, M. A., Caulfield, T. J., Kowala, M. C., and Valentine, M. T. (1999) Prospective identification of biologically active structures by topomer shape similarity searching. *J Med Chem* **42**, 3919–33.
67. Rarey, M. and Stahl, M. (2001) Similarity searching in large combinatorial chemistry spaces. *J Comput Aided Mol Des* **15**, 497–520.
68. Rarey, M. and Dixon, S. (1998) Feature trees: a new molecular similarity measure based on tree matching. *J Comput Aided Mol Des* **12**, 471–90.
69. Boehm, M., Wu, T. Y., Claussen, H., and Lemmen, C. (2008) Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J Med Chem* **51**, 2468–80.
70. Lessel, U., Wellenzohn, B., Lilienthal, M., and Claussen, H. (2009) Searching fragment spaces with feature trees. *J Chem Inf Model* **49**, 270–9.
71. Agarwal, A. K., Johnson, A. P., and Fishwick, C. W. (2008) Synthesis of de novo designed small-molecule inhibitors of bacterial RNA polymerase. *Tetrahedron* **64**, 10049–54.
72. Sova, M., Cadez, G., Turk, S., Majce, V., Polanc, S., Batson, S., Lloyd, A. J., Roper, D. I., Fishwick, C. W., and Gobec, S. (2009) Design and synthesis of new hydroxyethylamines as inhibitors of D-alanyl-D-lactate ligase (VanA) and D-alanyl-D-alanine ligase (DdlB). *Bioorg Med Chem Lett* **19**, 1376–9.
73. Park, H., Bahn, Y. J., and Ryu, S. E. (2009) Structure-based de novo design and biochemical evaluation of novel Cdc25 phosphatase inhibitors. *Bioorg Med Chem Lett* **19**, 4330–4.
74. Wang, R., Gao, Y., and Lai, L. (2000) LigBuilder: a multi-purpose program for structure-based drug design. *J Mol Model* **6**, 498–516.
75. Schüller, A., Suhartono, M., Fechner, U., Tanrikulu, Y., Breitung, S., Scheffer, U., Göbel, M. W., and Schneider, G. (2008) The concept of template-based de novo design from drug-derived molecular fragments and its application to TAR RNA. *J Comput Aided Mol Des* **22**, 59–68.
76. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-Hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed* **38**, 2894–6.
77. Mauser, H. and Guba, W. (2008) Recent developments in de novo design and scaffold hopping. *Curr Opin Drug Discovery Dev* **11**, 365–74.
78. Todorov, N. P., Buenemann, C. L., and Alberts, I. L. (2006) De novo ligand design to an ensemble of protein structures. *Proteins: Struct, Funct, Bioinf* **64**, 43–59.
79. Alberts, I. L., Todorov, N. P., and Dean, P. M. (2005) Receptor flexibility in de novo ligand design and docking. *J Med Chem* **48**, 6585–96.
80. Maass, P., Schulz-Gasch, T., Stahl, M., and Rarey, M. (2007) Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J Chem Inf Model* **47**, 390–9.
81. Grabowski, K., Proschak, E., Baringhaus, K., Rau, O., Schubert-Zsilaveczc, M., and Schneider, G. (2008) Bioisosteric replacement of molecular scaffolds: fromnatural products to synthetic compounds. *Nat Prod Commun* **3**, 1355–60.
82. Hartenfeller, M., Proschak, E., Schüller, A., and Schneider, G. (2008) Concept of combinatorial de novo design of drug-like molecules by particle swarm optimization. *Chem Biol Drug Des* **72**, 16–26.

Chapter 13

Classification of Chemical Reactions and Chemoinformatic Processing of Enzymatic Transformations

Diogo A.R.S. Latino and João Aires-de-Sousa

Abstract

The automatic perception of chemical similarities between chemical reactions is required for a variety of applications in chemistry and connected fields, namely with databases of metabolic reactions. Classification of enzymatic reactions is required, e.g., for genome-scale reconstruction (or comparison) of metabolic pathways, computer-aided validation of classification systems, or comparison of enzymatic mechanisms. This chapter presents different current approaches for the representation of chemical reactions enabling automatic reaction classification. Representations based on the encoding of the reaction center are illustrated, which use physicochemical features, Reaction Classification (RC) numbers, or Condensed Reaction Graphs (CRG). Representation of differences between the structures of products and reactants include reaction signatures, fingerprint differences, and the MOLMAP approach. The approaches are illustrated with applications to real datasets.

Key words: Chemical reactions, Enzymatic reactions, Metabolism, Self-organizing maps, Fingerprints, Reaction center, Molecular descriptors, Reaction descriptors

1. Introduction

The automatic processing of chemical reaction data has been at the core of chemoinformatics since its foundation [1] but has gained a particular interest in recent years due to the importance of metabolism and the emergence of large available databases of enzymatic reactions. These are providing an increasingly detailed picture of the cells' biochemical machinery and the associated genes [2–7].

Automatic classification of chemical reactions is of high importance for the analysis of reaction databases, reaction retrieval, reaction prediction, or synthesis planning. In bioinformatics, the automatic perception of chemical similarities between metabolic reactions is required for a variety of applications ranging

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_13, © Springer Science+Business Media, LLC 2011

from the computer-aided validation of classification systems to genome-scale reconstruction (or alignment) of metabolic pathways, the classification of enzymatic mechanisms, or enzymatic structure–function studies.

This chapter presents chemoinformatic approaches to reaction representation and classification and the diversity of their possible applications. In the selection of approaches, emphasis was placed on concepts with applications, particularly recent applications to datasets of metabolic reactions.

2. Representation and Classification of Reaction Centers

Chemical reactions involve the transformation of molecular structures at certain atoms and bonds, generally substructures of the reactants. The reaction center consists of the atoms and bonds directly involved in the reaction, e.g., bonds broken, made or changed, and corresponding atoms.

Reactions can be represented by descriptors of the reaction center similar to the representation of molecular structures by molecular descriptors. By focusing on the reaction center, these descriptors can capture similarities between similar reactions (or the same reaction) occurring with different substrates.

The calculation of descriptors for a reaction center requires the previous assignment of the reaction center, and in some cases atom-to-atom mapping, i.e., the correspondence between each atom in the reactants and each atom in the products – Fig. 1. Manual assignment of reaction centers is a tedious task, but it has been performed for databases of reactions. A few commercial software packages claim to automatically assign reaction centers from the structures of the reactants and products. Recently Körner and Apostolakis [8] proposed an imaginary transition state energy approach to reaction mapping and validated their algorithm with the manually annotated assignments of the 1,542 reactions in the BioPath database [9] – the algorithm found a correct mapping for >99% of the cases.

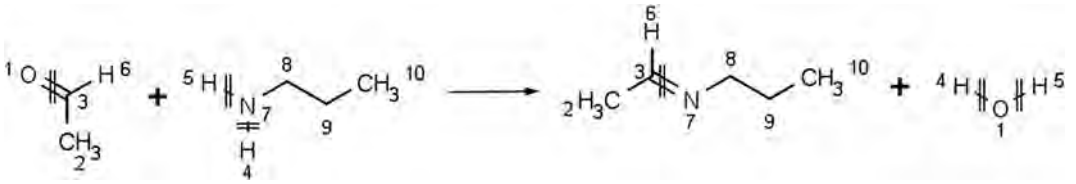


Fig. 1. Atom-to-atom mapping in a chemical reaction and identification of bonds involved in the reaction.

2.1. Representations Based on Connectivity Features

2.1.1. Reaction Classification Numbers

Kanehisa and coworkers developed the reaction classification (RC) numbers [10] for integrating chemical information in the KEGG database (Kyoto Encyclopedia of Genes and Genomes). RC numbers enable the representation and classification of the metabolic reactions in the database. First a reaction is decomposed into reactant pairs – a reactant and the product in which it is transformed. Then each pair is structurally aligned, i.e., a correspondence is established through graph matching between atoms of the reactant and the product, to identify molecular regions that match (M) or differ (D). The R-atom (for *R*eaction center) is defined as a matched atom that is adjacent to a non-matched atom. Non-matched atoms that are adjacent to the R-atom are referred to as D-atoms while matched adjacent atoms are referred to as M-atoms. When no difference region is found, i.e., the connectivity between all *non-hydrogen* atoms is the same, R-, D-, and M-atoms are defined in terms of changes in atom types.

The method uses a list of 68 predefined atom types such as the 23 atom types of carbon shown in Fig. 2. From this list, numerical codes are defined for conversion patterns of atom types, e.g., C1a → C1a; C1a → C1b; C1d + C5a → C5a + C6a. The RC numbers consist in the numerical codes for the conversion patterns in the three regions (R, D, and M). Figure 3 shows an example of the RC number generation for an enzymatic reaction.

The representation of chemical reactions by RC numbers has been applied to the automatic assignment of the Enzyme Commission (EC) numbers to enzymatic reactions. A query reaction is classified based on reactions sharing the same RC number in the KEGG database. Different restrictions are possible (e.g., R, D, and M patterns, or only R and D). This approach, which is accessible as a web service at http://www.genome.jp/ligand-bin/predict_reaction (accessed September 2009). In leave-one-out experiments, the number of hit reactions with the same and different EC sub-subclasses to the query was counted [10] obtaining about 90% accuracy (coverage: 62% of the dataset).

functional group	atom type	definition	functional group	atom type	definition
		Carbon (23 types)			
alkane	C1a	R–CH_3	alkyne	C3a	R≡C–H
	C1b	R–CH_2–R		C3b	R≡C–R
	C1c	R–CH(–R)–R	aldehyde	C4a	R–CH=O
	C1d	R–C(–R)$_2$–R	ketone	C5a	R–C(=O)–R
cyclic alkane	C1x	ring–CH_2–ring	cyclic ketone	C5x	ring–C(=O)–ring
	C1y	ring–CH(–R)–ring	carboxylic acid	C6a	R–C(=O)–OH
	C1z	ring–C(–R)$_2$–ring	carboxylic ester	C7a	R–C(=O)–O–R
alkene	C2a	R=CH_2		C7x	ring–C(=O)–O–ring
	C2b	R=CH–R	aromatic ring	C8x	ring–CH=ring
	C2c	R=C(–R)$_2$		C8y	ring–C(–R)=ring
cyclic alkene	C2x	ring–CH=ring	undefined C	C0	
	C2y	ring–C(–R)=ring			

Fig. 2. Definition of 23 KEGG atom types of carbon used for the generation of RC numbers. Adapted with permission from Kotera et al. [10]. © 2004 American Chemical Society.

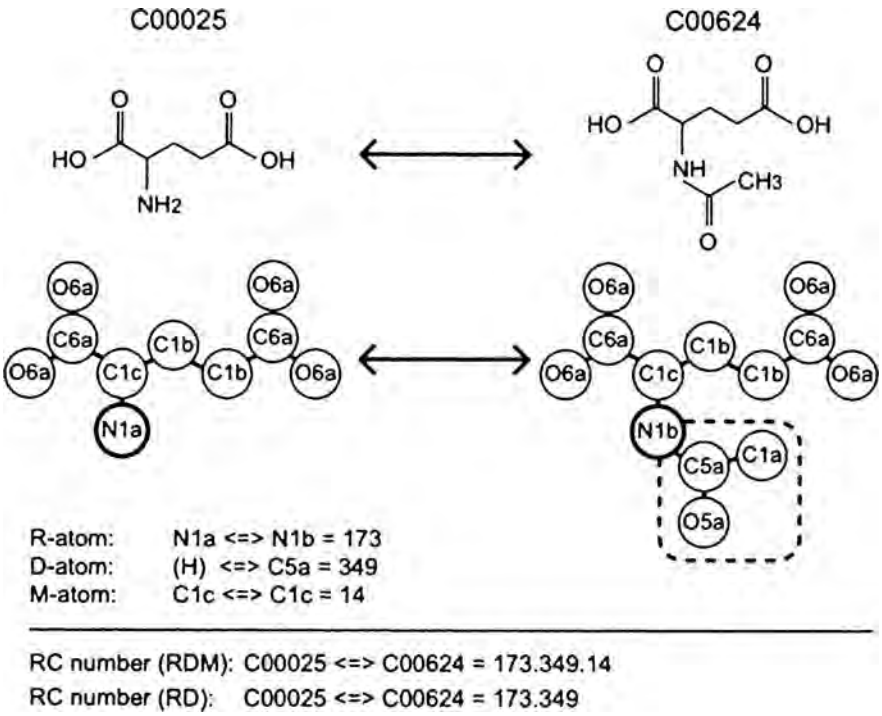


Fig. 3. Assignment of the RC number, which describes the conversion patterns of the KEGG atom types for the reaction center atom (R-atom), the difference structure atom (D-atom), and the matched structure atom (M-atom). Adapted with permission from Kotera et al. [10]. © 2004 American Chemical Society.

2.1.2. Fingerprints of Enzymatic Reaction Features

Focusing on reaction mechanisms, Mitchell and coworkers developed a reaction representation scheme for enzymatic reactions [11] that provides a quantitative measure of the similarity of reactions based upon their explicit mechanisms. Differently from other methods that encode the overall reaction, in this case the individual steps of the mechanism are explicitly represented. The authors aimed at a similarity measure *for enzymes* based on reaction mechanisms, motivated by cited evidence from the literature [12–14] suggesting that reaction steps are often conserved during the evolution of enzyme function.

To measure the similarity between individual steps of mechanisms, two methods were proposed: a simple bond change (BC) method encoding the bond changes that occur on going from the reactants to the products in each step of the mechanism – bond formations, bond cleavages, increases in bond order, and decreases in bond order; and a fingerprint (FP) method incorporating more detailed information about each mechanistic step.

For the FP method, 58 features of a reaction were used as reaction descriptors including the number of reactants, the difference between the number of products and reactants, the difference between the number of cycles in products and reactants, the number of times a bond type is involved in the reaction (from a list

of 21 bond types), the presence of cofactors, the number of occurrences of five types of bond order change (bond formation, bond cleavage, changes in order from 1 to 2, 2 to 1, and 3 to 2), charge changes by atom type, and involvement of radicals.

These fingerprints of enzymatic reaction features were applied to the assessment of similarity between individual steps of enzymatic reaction mechanisms. The quantitative measurement of the similarity between two enzymatic reactions was derived after the alignment of the individual steps of the two mechanisms. The study used 100 reaction mechanisms (395 reaction steps) taken from the MACiE database of enzyme reaction mechanisms [15] and identified some examples of convergent evolution of chemical mechanisms, i.e., cases where the same mechanisms are used by enzymes unrelated in sequence or structure.

2.1.3. Condensed Reaction Graphs

To generate a Condensed Reaction Graph (CRG), the reactants and products of a reaction are merged in an imaginary transition state, or pseudo-compound – Fig. 4. The reaction descriptors consist in the number of occurrences of predefined fragments in the pseudo-compound. Fragments include at least one bond of the reaction center and are predefined-sized sequences of atoms of specific atom types connected by bonds of specific bond types. Bond types are defined according to their fate in the reaction, e.g., "no bond" to single, single to double, double to "no bond."

CRG descriptors have been used for the assessment of similarity between reactions in datasets and can be applied to QSPR studies to predict properties of reactions, e.g., reaction rates [16, 17].

2.2. Representations Based on Physicochemical Parameters

Physicochemical and energy parameters calculated for atoms and bonds of the reaction center (or its neighbors) provide powerful ways to represent and classify chemical reactions. They enable a high level of abstraction in the comparison of reactions, revealing intrinsic similarities between structurally diverse reactions. Furthermore, as reaction mechanisms are usually not reported in most current reaction databases or are even unknown for large numbers of reactions, considering physicochemical parameters of reactants and products is an indirect way of taking mechanistic aspects into account.

Significant contributions to this field came from the research groups of Gasteiger and Funatsu during the 1980s and 1990s.

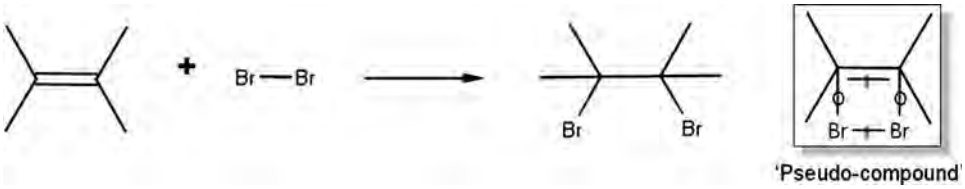


Fig. 4. The CRG approach is based on an imaginary transition state (or pseudo-compound) with bond types defined according to their fate in the reaction.

A bold statement can be cited from a 1994 paper of Rose and Gasteiger [18]: "We strongly believe that any progressive scheme for the classification of chemical reactions should be able to uncover and incorporate the physicochemical driving forces of chemical reactions, since the electronic, energy, and steric effects determine a reaction mechanism and the reaction mechanism determines the reaction conditions. Clearly, common reaction mechanisms would form the soundest foundation for reaction classification."

In order to process large databases of chemical reactions, fast methods are required to calculate electronic factors such as charge distribution and inductive and resonance effects at the reaction center. Rapid empirical methods have been implemented for the calculation of atomic charges, parameters for the inductive, resonance, and polarizability effects, and bond dissociation energies. The calculation of atomic charges can be performed with the partial equalization of orbital electronegativity (PEOE) algorithm [19]. The calculation of the π-atomic charges and π-electronegativities is obtained through generation and weighting of the various resonance structures of a molecule. The resonance parameters quantify the stabilization of a positive or negative charge obtained in the formal polar breaking of a bond. Bond polarizabilities and bond dissociation energies are obtained by additive schemes.

In a set of reactions with a common reaction center, parameters can be defined for specific atoms or bonds of the reaction center, yielding a fixed number of descriptors for each reaction – a vector. This approach is illustrated with the classification of a dataset comprising the 120 reactions from the 1992 edition of the ChemInform RX database that involve the addition of an H–C bond to a C=C bond to form a new C–C bond [20]:

$$C(1) = C(2) + H - C(3) \rightarrow H - C(1) - C(2) - C(3)$$

Seven properties were chosen on the basis of chemical intuition to encode the reactions: σ- and π-electronegativities at atoms C(1) and C(3), total charges on atoms C(2) and C(3), and effective polarizability on atom C(3).

The 120 reactions (encoded by the seven features) were mapped on a Kohonen neural network (or self-organizing map, SOM). This is a 2D grid of so-called neurons that distribute objects (reactions in this case) according to similarities between their features (physicochemical parameters in this case). At the end of the training, each neuron is inspected for the objects activating it and is classified according to the classes of these objects. Because only the features of the objects influence the mapping (not their classes), the training of a SOM is an unsupervised process. The resulting map for the dataset of 120 reactions is shown in Fig. 5. Classes of the reactions were manually assigned by chemists in order to assess the quality of the mapping. The figure clearly

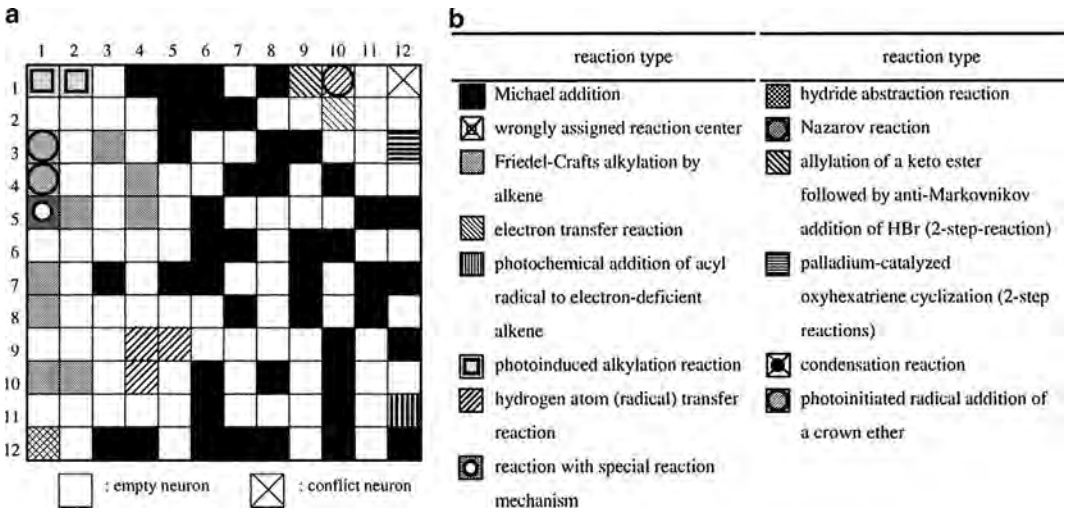


Fig. 5. Kohonen map obtained for the classification of 120 reactions with the occupied neurons marked with different symbols according to intellectually assigned reaction types. Adapted with permission from Chen and Gasteiger [20]. © 1997 American Chemical Society.

reveals clustering according to reaction types, and the authors reported that in all cases but one the reactions activating a neuron were from the same type. The only conflict corresponded to a reaction with a wrongly assigned reaction center in the database. The model was further validated with external datasets yielding correct predictions for up to 86% of the cases.

A similar methodology could be applied to a dataset of reactions not exactly with the same reaction center, but having in common an oxygen atom at the reaction center [21]. Satoh et al. represented the reactions by the changes in physicochemical properties at the oxygen atoms of the reaction sites – the properties of the oxygen atom in the product minus the properties in the reactant. The chosen parameters were the σ- and π-charge, σ- and π-electronegativity, polarizability, and pK_a values. Figure 6 illustrates the encoding of a reaction. This representation was used to train a Kohonen SOM with 152 O-atoms from 131 reactions obtaining the map of Fig. 7.

The SOM placed reactions belonging to similar reaction types on the same or neighboring neurons, which indicate that changes in the electronic properties of the reaction centers correspond quite well with changes in the substructures at the reaction sites and also with reaction types and categories manually assigned by chemists. The authors concluded that this representation of reactions is a valuable approach to the identification of reaction types and categories that have been established by chemists through analyses of reactions from several different points of view.

More recently, 135 enzymatic hydrolysis reactions falling into the domain of the EC class three were compared with

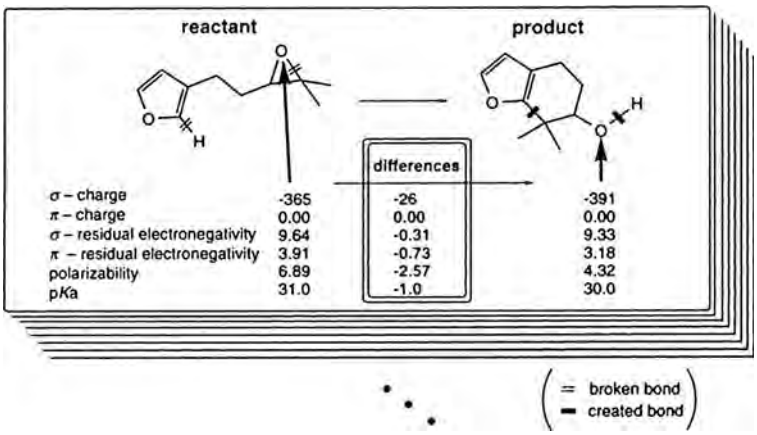


Fig. 6. The changes in σ-charge, π-charge, σ-residual electronegativity, π-residual electronegativity, polarizability, and pK_a values at oxygen atoms of the reaction sites in going from the reactants to the products were taken as a characterization of the individual reactions and were used for their classifications. In this example, differences in these values between an oxygen atom in the epoxide of the reactant to an oxygen atom in the hydroxy group of the product are calculated. Reprinted with permission from Satoh et al. [21]. © 1998 American Chemical Society.

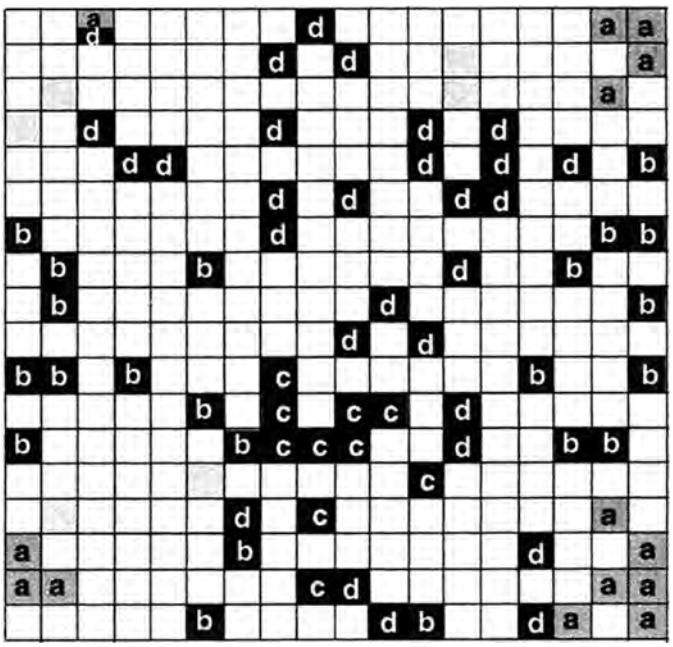


Fig. 7. Distribution of 152 O-atoms from 131 reactions on a Kohonen SOM. These are classified into four reaction types: reductions and alkylations (labeled a); cleavage of epoxides, ethers, lactones, and esters (labeled b); oxidation of alcohols (labeled c); and formation of epoxides, ethers, lactones, and esters (labeled d). Reprinted with permission from Satoh et al. [21]. © 1998 American Chemical Society.

physicochemical properties of the reaction center, namely of bonds reacting in the substrate [22]. The authors chose these reactions as they have similar overall reaction patterns, and the EC classification system within the class is largely built on criteria based on the reaction mechanism. Six properties were calculated for the bonds reacting in the substrate: difference in partial atomic charges between the two atoms of the bond, difference in σ-electronegativities, difference in π-electronegativities, effective bond polarizability, delocalization stabilization of a negative charge generated by heterolysis of the bond, and delocalization stabilization of a positive charge. Each bond is thus characterized by six descriptors. Most of the reactions in the dataset have only one reacting bond in the substrate, but some have two reacting bonds. In order to represent all reactions by the same number of descriptors, 12-dimensional feature vectors were used – vectors of reactions with only one reacting bond had six coordinates filled with zero values.

Mapping the 135 reactions on a SOM showed similarities between reactions that reproduce well the classification of enzymes by the EC number. Inspection of the substrates revealed correlations between substructures and the distribution of reactions in some regions. Furthermore, the map showed how the structural environment of a reacting bond influences its physicochemical parameters.

A similar approach was followed to map a larger set of reactions – 33,613 reactions from the Theilheimer database [23, 24] were encoded with six physicochemical properties for the bonds of products at the reaction center. The dataset was highly heterogeneous concerning the reaction center, and the reactions were represented by vectors with room for up to six bonds (6 bonds × 6 properties = 36 features). The bonds of the reaction center were ranked for the assignment of positions in the vector. The 92 × 92 Kohonen SOM trained with the database showed reasonable grouping according to structural features of the reaction center and enabled the comparison of two large databases in terms of their diversity (by mapping both on the same SOM). Other databases, with reactions tagged with a date, were mapped on the reference SOM. The differences observed for subsets of different years revealed, for some types of reactions, time trends in the number of reactions abstracted from the literature [23].

3. Representation of Differences Between the Structures of Products and Reactants

The representation of chemical reactions does not necessarily require the extraction and encoding of explicit information about the reaction center. When reactions are stoichiometrically balanced, reaction descriptors can be derived from the comparison

of *molecular* descriptors of products and reactants. As a reaction transforms molecular structures of reactants into products, the overall transformation can be captured by the changes in their molecular descriptors. Two different approaches will be separately explained, one based on molecular fingerprints or signatures, and the other on MOLMAP physicochemical/topological descriptors of the types of bonds available in a molecule.

3.1. Differences of Molecular Fingerprints or Signatures

A method implemented in Daylight software can represent reactions by "difference Daylight fingerprints." These are obtained from the difference between Daylight fingerprints of reactant and product molecules and the fingerprints of product molecules. [25] As the molecular fingerprints encode the presence of fragments (paths), the difference in the fingerprint of the reactant molecules and the fingerprint of the product molecules reflects the bond modifications caused by the reaction – the paths appearing both in the reactants and products correspond to unchanged substructures and the fingerprint subtraction cancels them. But fingerprints are binary, and multiple occurrences of a path are not encoded; therefore, a simple subtraction is not enough. It is required to keep track of the count of each path in the reactant and product and then subtract the counts of a given path. If the difference in count is nonzero, then the path is used to set a bit in the difference fingerprint. If the difference in count is zero, then no bit is set for that path in the difference fingerprint. This method avoids assignment of reaction centers and atom-to-atom mapping.

Ridder and Wagener [26] implemented a similar idea but relied on Sybyl atom types and atom types augmented with a single layer around the central atom. The difference fingerprint was defined as the differences in occurrence of each atom type in the reactant and product fingerprints. The method was applied to the classification of metabolic reactions to assist in the establishment of rules for reaction prediction. The reactions of a training set were projected on a 2D plane to optimally reflect reaction fingerprint distances calculated between all pairs of reactions. The method is based on stochastic proximity embedding and optimizes the distances between points on a 2D plane to correspond as much as possible to the distances calculated in the fingerprint space between all pairs of metabolic reactions.

Instead of fingerprints, Faulon et al. [27, 28] explored differences in "molecular signatures" to represent chemical reactions. Molecular signatures are made of atomic signatures, and an atomic signature is a canonical representation of the subgraph surrounding a particular atom. The subgraph includes all atoms and bonds up to a predefined distance from the given atom (the height, h). Each component of a molecular signature counts the number of occurrences of a particular atomic signature in the molecule. "Reaction signatures" are then defined as the difference between the molecular

signatures of the products and the molecular signatures of the substrates. This method was applied to the automatic assignment of EC numbers in a dataset of 6,556 reactions from the KEGG database using support vector machines (SVM). Correct assignment of EC class, subclass, and sub-subclass in cross-validation experiments was reported in up to 91, 84, and 88% respectively.

3.2. The MOLMAP Approach

Physicochemical properties of atoms and bonds at the reaction center were shown to provide a chemically meaningful way of comparing and classifying reactions into reaction types (Sect. 2.2).

In order to incorporate the information concerning bond properties for an entire molecule, and at the same time having a fixed-length descriptor, independent of the molecular size, *MOLMAP* representations (*MOL*ecular *M*ap of *A*tom-level *P*roperties) were put forward that map all the bonds of a molecule into a fixed-length 2D SOM [29].

A SOM must be trained beforehand with a diversity of bonds from different structures (each bond described by bond properties, e.g., physicochemical properties calculated by the above-mentioned PEOE method). Then all the bonds of one molecule are submitted to the trained SOM, each bond activates one neuron, and the pattern of activated neurons represents the bonds available in that structure. The activated neuron is the neuron with the minimum Euclidean distance to the bond features vector. By counting the number of bonds that activate each neuron, a numerical descriptor is obtained – the MOLMAP. The size of the MOLMAP, i.e., the number of its components, corresponds to the number of neurons in the SOM. To focus on functional groups, MOLMAPs may be restricted to bonds involving (or at a one bond distance of) specific types of atoms (e.g., heteroatoms or atoms belonging to a π system). Additionally, the MOLMAP component corresponding to a neuron may take into account the frequency of activation of neighbor neurons. The MOLMAP descriptors can be directly used for data mining or QSAR studies related to chemical reactivity, in situations involving different types of reaction sites in a single dataset, more than one reaction site in a single structure, or unknown reaction sites [29–32].

For a stoichiometrically balanced reaction, the difference between the MOLMAPs of the products and the MOLMAPs of the reactants can be interpreted as a MOLMAP representation *of the reaction* – Fig. 8. The MOLMAP of a molecule is the pattern of neurons that are activated by the bonds existing in that molecule. Bonds far apart from the reacting center are mostly unchanged during the reaction, exhibiting almost the same physicochemical properties, and thus activating the same position (neuron) of the MOLMAP both in the reactants and in the products. Therefore, the difference map (MOLMAP of the reaction) gets null contributions from them. The pattern of neurons in the

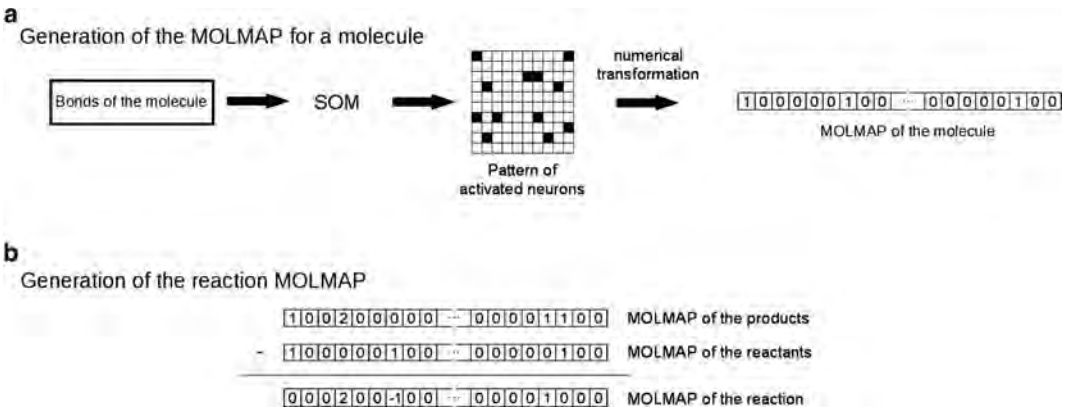


Fig. 8. Procedure for the generation of reaction MOLMAPs.

MOLMAP of the reaction with nonzero values relates to the bonds of the reactants that break or change properties and to the bonds of the products that are made or changed in the reaction. The former leads to negative values, while the latter leads to positive values.

MOLMAPs were applied to the classification and data mining of >3,000 enzymatic reactions from the KEGG database (the so-called *reactome*) [33–35]. In order to investigate the correlation between similarities in reaction MOLMAPs and similarities in EC numbers, the reactions were submitted to a SOM. It must be emphasized that this SOM is independent of the SOM defined for the generation of MOLMAPs. While the objects presented to the former are bonds, the objects presented to the new SOM are reactions. The neurons of the new SOM were colored after the training, according to the classes of the reactions that activated them. The resulting 49 × 49 Kohonen SOM for an experiment with 7,482 enzymatic reactions (all original reactions were included in both directions) shows a remarkable clustering of the reactions according to the first digit of the EC classification (Fig. 9). The three most populated classes showed the best clustering – oxidoreductases, transferases, and hydrolases. Ligases also exhibited good clustering. Reactions were encoded by MOLMAPs of size 625 (25 × 25) generated from topological and physicochemical bond descriptors obtained with ChemAxon software [36].

After the neurons of a SOM are classified ("colored") on the basis of a training set of reactions, it can be applied to classify new reactions. Experiments were performed with reactions from all EC classes in one SOM to assign the first digit of the EC number. Separate SOMs were trained for different classes to assign the subclass and sub-subclass – also at these levels, the results of unsupervised mapping showed a reasonable agreement with the EC classification. The assignment of EC numbers at the class, subclass, and sub-subclass levels in independent test sets

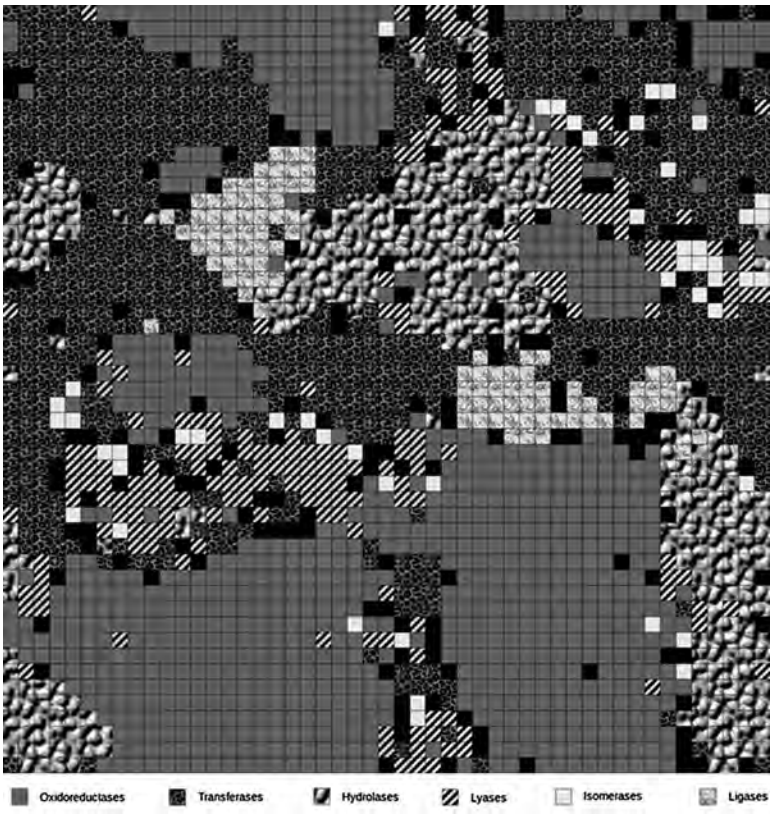


Fig. 9. Surface of a 49 × 49 Kohonen SOM with toroidal topology trained with 7,482 enzymatic reactions encoded by MOLMAPs of size 625 using topological and physicochemical descriptors. Each neuron was classified after the training, according to the reactions that were mapped onto it or onto its neighbors. *Black* neurons correspond to conflicts.

was accomplished with accuracies up to 92, 80, and 70%, respectively. These results were obtained with consensus predictions from ten independently trained SOMs for each task, in order to overcome fluctuations caused by the random factors included in the SOM algorithm. The procedure was also evaluated with increasingly different training and test sets. For example, in the most stringent experiment, all the reactions of the test set belonged to different sub-subclasses of those in the training set, and the class was correctly predicted for 68% of the test set. This result gives an indication of the similarity of reactions across sub-subclasses (within the same class). At the same time, experiments to predict the third digit of the EC number demonstrated the ability of MOLMAP descriptors to discriminate sub-subclasses.

SOMs trained with reactions from all classes were used to data mine the database in order to detect very similar reactions with differences in the EC number at the class level. In one case, [33] with MOLMAPs generated from seven physicochemical

properties calculated by *PETRA* [37], a reaction tagged with a lyase (EC 4.2.99.16) was identified on a neuron surrounded by reactions catalyzed by transferases, and striking similarities were indeed observed between the reactions. The research of the literature [38] revealed that the lyase had its EC number changed into a new number with the same first three digits as the reactions in the neighbor neurons! In another case [34], the neurons of a SOM were screened to find reactions of different classes in the same neuron, and a list was produced with pairs of very similar reactions tagged with different EC classes. It was suggested that some of the detected enzymes might deserve a revision of EC numbers while others illustrated problematic aspects of the application of EC rules related to reversibility of reactions, or even enzymes catalyzing more than one type of reaction but getting a single EC number. Some of the identified pairs illustrated limitations of the MOLMAP approach in processing highly diverse datasets. Examples include some reactions with different types of bonds broken in the reactants that were perceived as similar because of the formation of similar bonds in the products, and reactions in which reactants very similar to the products result in MOLMAPs with only few non-null components (the reactions may yield globally similar almost-null MOLMAPs although the non-null values are different because they correspond to different types of bonds) [34].

Training of *Random Forests* (a supervised machine-learning algorithm) with the same datasets of enzymatic reactions encoded by the same MOLMAP descriptors to automatically assign EC numbers resulted in more accurate predictions than those obtained with SOMs (an unsupervised method). The accuracies reached 95, 90, and 85% for the class, subclass, and sub-subclass level if examples of reactions with the same EC number were allowed in the training and test sets simultaneously. In the absence of reactions belonging to the same sub-subclass (lower similarity between training and test sets), accuracies dropped to 72% in the prediction of the class [35].

4. Conclusions

Several options are currently available for the representation of chemical reactions, their clustering and automatic classification. Although more specific information is used by methods explicitly encoding features of the reaction center, "difference methods" only based on the structures of reactants and products (not requiring the assignment of reaction centers) were shown to perform very well in various applications. When processing datasets with no common reacting substructures, the representation of reaction

centers requires a scheme to align atoms and/or bonds, and to process reactions with different numbers of bonds involved.

Representations based on topological features or atom types may be easy to interpret, and conservative at detecting similarities between reactions, but they may lack the ability to detect similar electronic effects caused by different types of atoms and thus to generalize types of reactions. Physicochemical parameters can capture reaction mechanism information on a more general level. The increasing accessibility of quantum calculations at higher levels of theory can possibly bring exciting developments to the classification of chemical reactions based on physicochemical and energy parameters.

Acknowledgments

Diogo A. R. S. Latino acknowledges Fundação para a Ciência e Tecnologia (Ministério da Ciência, Tecnologia e Ensino Superior, Lisbon, Portugal) for financial support under Ph.D. grant SFRH/BD/18347.

References

1. Chen, L. (2003) Reaction classification and knowledge acquisition, In *Handbook of chemoinformatics: from data to knowledge*. Gasteiger, J. and Engel, T. (Eds.). Wiley-VCH, New York, Vol. 1, pp 348–388.
2. Goto, S., Nishioka, T., and Kanehisa, M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics* **14,** 591–599.
3. Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30.
4. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32,** D277-D280.
5. Kanehisa, M. (1997) A database for post-genome analysis. *Trends Genet.* **13,** 375–376.
6. Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., Walk, T. C., Zhang, P., and Karp, P. D. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc ollection of Pathway/Genome Databases. *Nucleic Acids Res.* **36,** D623-D631.
7. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37,** D619-D622.
8. Körner, R. and Apostolakis, J. (2008) Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J. Chem. Inf. Model.* **48,** 1181–1189.
9. Apostolakis, J., Sacher, O., Körner, R., and Gasteiger, J. (2008) Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J. Chem. Inf. Model.* **48,** 1190–1198.
10. Kotera, M., Okuno, Y., Hattori, M., Goto, S., and Kanehisa, M. (2004) Computational assignement of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **126,** 16487–16498.
11. O'Boyle, N. M., Holliday, G. L., Almonacid, D. E., and Mitchell, J. B. O. (2007) Using reaction mechanism to measure enzyme similarity. *J. Mol. Biol.* **368,** 1484–1499.
12. Babbitt, P. C. and Gerlt, J. A. (1997). Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution

of new catalytic activities. *J. Biol. Chem.* **272,** 30591–30594.
13. Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307,** 1113–1143.
14. Bartlett, G. J., Borkakoti, N., and Thornton, J. M. (2003). Catalysing new reactions during evolution: economy of residues and mechanism. *J. Mol. Biol.* **331,** 829–860.
15. Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle, N. M., Murray-Rust, P., Thornton, J. M., and Mitchell, J. B. O. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics 21,* 4315–4316.
16. Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V. P. (2005) Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **19,** 693–703.
17. Fujita, S. (1986) Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Comput. Sci.* **26,** 205–212.
18. Rose, J. R. and Gasteiger, J. (1994) HORACE: an automatic system for the hierarchical classification of chemical reactions. *J. Chem. Inf. Comput. Sci.* **34,** 74–90.
19. Gasteiger, J. and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity – A rapid access to atomic charges. *Tetrahedron* **36,** 3219–3228.
20. Chen, L. and Gasteiger, J. (1997) Knowledge discovery in reaction databases: landscaping organic reactions by a self-organizing neural network. *J. Am. Chem. Soc.* **119,** 4033–4042.
21. Satoh, H., Sacher, O., Nakata, T., Chen, L., Gasteiger, J., and Funatsu, K. (1998) Classification of organic reactions: similarity of reactions based on changes in the electronic features of oxygen atoms at the reaction sites. *J. Chem. Inf. Comput. Sci.* **38,** 210–219.
22. Sacher, O., Reitz, M., and Gasteiger, J. (2009) Investigations of enzyme-catalyzed reactions based on physicochemical descriptors applied to hydrolases. *J. Chem. Inf. Model.* **49,** 1525–1534.
23. Sacher, O. (2001) Classification of Organic Reactions by Neural Networks for the Application in Reaction Prediction and Synthesis Design. Ph.D. Thesis, University of Erlangen-Nuremberg, Erlangen, Germany, http://www2.chemie.uni-erlangen.de/services/dissonline/data/dissertation/Oliver_Sacher/html/ (accessed September 2009).
24. In 2000 the Theilheimer database was developed by MDL Information Systems, Inc., San Leandro, CA, USA.
25. Daylight (2008) Daylight Theory Manual, Daylight version 4.9, release date 02/01/08, Daylight Chemical Information Systems, Inc., http://www.daylight.com/dayhtml/doc/theory (accessed September 2009).
26. Ridder, L. and Wagener, M. (2008) SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* **3,** 821–832.
27. Faulon, J.-L., Visco, D. P., and Pophale, R. S. (2003) The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **43,** 707–720.
28. Faulon, J.-L., Misra, M., Martin, S., Sale, K., and Sapra, R. (2008) Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **24,** 225–233.
29. Zhang, Q.-Y. and Aires-de-Sousa, J. (2005) Structure-based classification of chemical reactions without assignment of reaction centers. *J. Chem. Inf. Model.* **45,** 1775–1783.
30. Gupta, S., Matthew, S., Abreu, P. M., and Aires-de-Sousa, J. (2006) QSAR analysis of phenolic antioxidants using MOLMAP descriptors of local properties. *Bioorg. Med. Chem.* **14,** 1199–1206.
31. Zhang, Q.-Y. and Aires-de-Sousa, J. (2007) Random forest prediction of mutagenicity from empirical physicochemical descriptors. *J. Chem. Inf. Model.* **47,** 1–8.
32. Carrera, G., Gupta, S., and Aires-de-Sousa, J. (2009) Machine learning of chemical reactivity from databases of organic reactions. *J. Comput. Aided Mol. Des. 23,* 419–429.
33. Latino, D. A. R. S. and Aires-de-Sousa, J. (2006) Genome-scale classification of metabolic reactions: a chemoinformatics approach. *Angew. Chem. Int. Ed.* **45,** 2066–2069.
34. Latino, D. A. R. S., Zhang, Q.-Y., and Aires-de-Sousa, J. (2008) Genome-scale classification of metabolic reactions and assignment of EC numbers with self-organizing maps. *Bioinformatics* **24,** 2236–2244.
35. Latino, D. A. R. S. and Aires-de-Sousa, J. (2009) Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests. *J. Chem. Inf. Model.* **49,** 1839–1846.
36. ChemAxon Kft., Budapest, Hungary, www.chemaxon.com Details about PETRA software are available from http://www2.chemie.uni-erlangen.de/software/petra (accessed September 2009).
37. PETRA is developed by Molecular Networks GmbH (Erlangen, Germany, http://www.mol-net.de).
38. http://www.chem.qmul.ac.uk/iubmb/enzyme/EC4/ 2/99/16.html accessed October 2009.

Chapter 14

Informatics Approach to the Rational Design of siRNA Libraries

Jerry O. Ebalunode, Charles Jagun, and Weifan Zheng

Abstract

This chapter surveys the literature for state-of-the-art methods for the rational design of siRNA libraries. It identifies and presents major milestones in the field of computational modeling of siRNA's gene silencing efficacy. Commonly used features of siRNAs are summarized along with major machine learning techniques employed to build the predictive models. It has also outlined several web-enabled siRNA design tools. To face the challenge of modeling and rational design of chemically modified siRNAs, it also proposes a new cheminformatics approach for the representation and characterization of siRNA molecules. Some preliminary results with this new approach are presented to demonstrate the promising potential of this method for the modeling of siRNA's efficacy. Together with novel delivery technologies and chemical modification techniques, rational siRNA design algorithms will ultimately contribute to chemical biology research and the efficient development of siRNA therapeutics.

Key words: siRNA, RNAi, Gene silencing, Bioinformatics, Cheminformatics, QSAR

1. Introduction

RNA interference (RNAi) is a posttranscriptional mechanism that modulates the expression of specific genes. It is mediated by double-stranded RNA (dsRNA) or short hairpin RNA (shRNA) molecules [1, 2]. As virtually every gene can be targeted via RNAi, there have been many efforts to develop technologies to harness this capability for both basic biomedical research and the development of potential RNAi therapeutics. As a result, synthetic short interfering RNA (or siRNA) technology has been developed that can systematically knock-down specific genes and explore the functional outcome. The siRNA technology has now become a standard tool for target identification and target validation in the pharmaceutical and biotechnology industry. siRNA molecules, coupled with novel delivery techniques [3] or chemical

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_14, © Springer Science+Business Media, LLC 2011

modification [4], have also been investigated as potential therapeutic agents for targeted cancer therapy or other diseases.

Figure 1a shows a schematic structure of a siRNA molecule, which is typically a 19–21-nucleotide (NT) double-stranded RNA, with 3′-end 2-NT overhangs. During the process of RNAi (Fig. 1b), one strand of the siRNA (called the guide strand) is incorporated into the RNA-induced silencing complex (RISC), and the other strand (called the passenger strand that should have the same sequence as the target mRNA) is degraded. The guide strand then base-pairs with a complementary target mRNA, inducing the cleavage of the mRNA molecule. As a result, it knocks down the expression of the corresponding gene.

It is known that only a fraction of siRNAs that satisfy Watson–Crick matching with the target mRNA are effective at silencing the expression of the target gene. Thus, it is highly desirable to develop effective rational selection criteria or models that can assist the design of siRNAs for RNAi experiments and the development of potential siRNA therapeutics.

The overall workflow (Fig. 2) for siRNA design involves (1) input target mRNA sequence; (2) generation of complementary sequences of siRNA guide (antisense) strands; (3) filtration of the

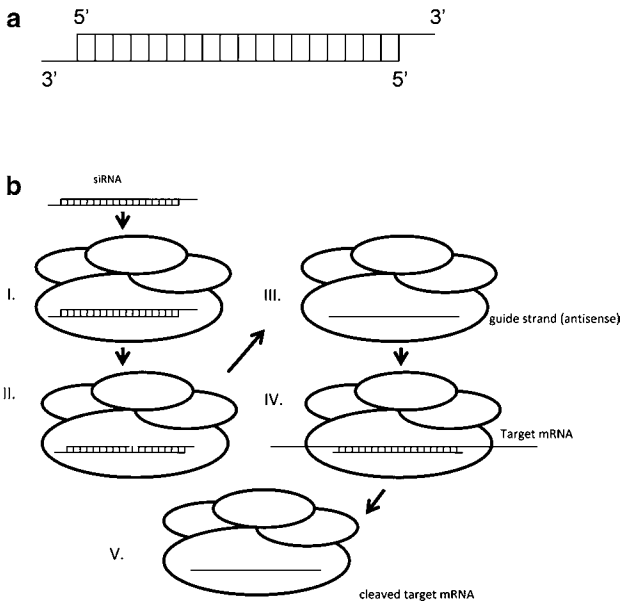


Fig. 1. (**a**) A siRNA molecule has two strands of oligonucleotide sequences. One strand is the guide strand (or antisense strand), and the other strand is the passenger strand (or sense strand). Each strand has a 2 NT (nucleotide) overhang at the 3′ end. (**b**) The mechanism of siRNA gene silencing is shown: first, double-stranded siRNA is bound to the RISC complex (I), then the passenger strand is cleaved (II) and released, leading to activated RISC with a bound guide strand (III). The guide strand directs the RISC complex to a complementary mRNA sequence (IV), and the mRNA is then cleaved and degraded (V).

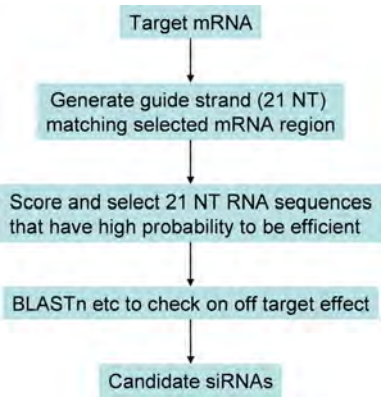


Fig. 2. General workflow for siRNA design and selection.

sequences based on unwanted motifs; (4) calculation of guide strand sequence and structural features; (5) scoring of siRNA potency based on either rule-based scoring functions or machine learning models; and (6) elimination of sequences with potential off target effect by sequence homology analysis using such tools as BLASTn.

Many aspects need to be considered in filtering siRNA sequences. Patzel [5] has summarized various factors that need to be considered depending on the intended use of the siRNAs. The factors include avoiding sequence motifs interfering with RNA synthesis and purification; modification of the siRNA molecules to include more lipophilic moieties to enhance the cellular delivery for in vivo use of siRNA; chemical modification to enhance siRNA's stability against nuclease-mediated degradation; avoidance of immunostimulatory sequence motifs; and thermodynamic accessibility of target mRNA structure as well as the avoidance of off target sequence match. Interested readers are referred to Patzel [5] for more details in regards to these aspects. Clearly, rational design of siRNA libraries is a multi-factorial problem.

This chapter, however, only focuses on the sequence-based predictive modeling of siRNA silencing potency, which is in the core of any computational tool for the rational design of siRNAs.

In the past 5 years, a variety of approaches to rational siRNA design have been reported. These publications have examined different features describing siRNA sequences and analyzed their relevancy to siRNA efficacy using different statistical and machine learning techniques. In this chapter, we aim to summarize some of these features, highlight several machine learning tools and review the guidelines and models developed thus far. We will also summarize a few online tools for the rational design of siRNAs. In addition, we will describe a new approach to siRNA efficacy modeling using ideas from QSAR (Quantitative Structure–Activity Relationship) studies. We hope to introduce this subject

to the researchers in the cheminformatics community so that they may investigate other novel ways to model gene silencing potency of siRNAs.

2. Computational Methods for siRNA Efficacy Prediction

2.1. Machine Learning and Statistical Methods

Many different machine learning algorithms have been used in the modeling of siRNA efficacy. The most popular ones are the Support Vector Machine (SVM) methods [6], Artificial Neural Network (ANN) methods [7], decision tree [8], and linear regression method [9]. Since most of these tools have been widely used in the cheminformatics community to build QSAR models, we will not detail these methods in this chapter. Interested readers are referred to the cited publications for more details.

2.2. Description of siRNAs

An important step in modeling siRNA gene silencing potency is to design calculable features that may be relevant to its efficacy. Most of the work published so far employs sequence based features that can be derived from siRNA molecules. These features include, but are not limited to, position-specific nucleotide preferences, overall nucleotide content and sequence motifs, thermodynamic properties, secondary structural features of both the siRNA and the target mRNA molecules, and other sequence derived features. We will summarize some of these features as follows.

2.2.1. Position-Specific Nucleotide Occurrences

As described above, siRNA molecules have two strands, the antisense strand (or guide strand) and the sense strand (or passenger strand). In modeling siRNA potency, one should choose consistently either the antisense or the sense strand as the basis for its analysis. Once the strand is chosen, one can study siRNAs by examining the identity of nucleotide at specific positions from 5′ to 3′ end. One can then examine the preferences of nucleotides at specific positions to see if there are significant differences for efficient and inefficient siRNAs in terms of position-specific nucleotide occurrences. If certain nucleotides are preferred for functionally efficient siRNAs, those preferences can be summarized into rules. This approach was very popular in early development of siRNA design rules. Position-specific nucleotide preferences are often interpreted as a result of the molecular mechanism of the RNAi process, and thus shed light on potential mechanisms of RNAi.

2.2.2. Nucleotide Content and Sequence Motifs

Many siRNA design rules examined overall content of nucleotide bases (G/C) in the siRNA sequences. Because the GC content has to do with the stability or secondary structure forming potentials of RNA molecules, it is thought to have an effect on the siRNA

efficacy. For example, Chan et al. [10] recently examined the effect of GC-content on the efficiency of RNAi. They have looked at the secondary structures for both the target mRNA and the siRNA guide strand. They found that the target site accessibility is more important than GC-content per se to siRNA efficiency, and that there is a high correlation between target site accessibility and the GC-content. Thus, %GC content is an indirect indicator of the accessibility of the target mRNA, thus contributes to siRNA efficacy. Percentage of GC content is also used in machine learning algorithms as part of the input descriptors.

Position independent sequence motifs of 2, 3, 4, or more bases were also found to be important for the discrimination of functional and nonfunctional siRNAs. Thus, the frequencies of all 2-base motifs, 3-base motifs, 4-base motifs, or even higher order motifs in siRNA sequences are often used as part of the input parameters for machine learning algorithms.

2.2.3. Thermodynamic Properties

The accessibilities of siRNA and target mRNA for hybridization, measured by folding free energy change, are shown to be significantly correlated with siRNA's gene silencing efficacy. Many algorithms take into account the thermodynamic profiles of siRNA duplexes or the hybridization thermodynamics between siRNA and the target mRNA. For example, Lu and Mathews [11] examined the equilibrium of siRNA-target hybridization and used it for the selection of efficient siRNAs. The predicted thermodynamic features are often used as part of the input for machine learning algorithms to model siRNA efficacy. Shabalina et al. [12] examined and optimized the free energy difference between the 5′ end and the 3′ end of the antisense strand, which is thought to be important for the discrimination between the two stands and thus contribute to siRNA efficacy. It has also been reported that potent siRNAs tend to have less stable 5′ end on the guide strand. For example, thermodynamic stability of the first four NT bases on the guide strand is a critical factor for siRNA efficacy [13].

2.3. State-of-the-Art Models of siRNA Efficacy Prediction

2.3.1. Rule-Based Methods

These methods aimed to derive simple rules thought to govern functional siRNAs [14, 15]. For example, GC content between 30 and 70% was considered to be good for functional siRNA, and a GC content of around 50% was considered as the most favorable. Others have looked at the nucleotide (NT) preferences at different positions. For example, "A/U" at the 5′ end of the antisense strand, and "G/C" at 5′ end of the sense strand were considered to be preferred. An "A" at position 3, a "U" at position 10 and ("A" or "U" or "C") at position 13 relative to the 5′ end of the sense strand were correlated to potency [5]. However, only limited consistency between the identified rules has been observed, probably due to the small heterogeneous datasets from which respective rules or guidelines were derived [16–18].

The guidelines proposed by Reynolds [16] and Ui-Tei [17] are two typical examples of the early development of rule based approaches to siRNA design. In the following, we describe in more detail the rules developed by Reynolds et al. [16] as an example of this type of approaches.

Reynolds et al. [16] tried to define patterns found in the functional siRNAs as the guidelines or criteria for rational design of siRNAs. They have examined 180 siRNAs targeting the mRNAs of two genes at every other position of the two 197-base regions of the firefly luciferase and human cyclophilin B mRNAs. Eight characteristics of the siRNAs were considered, including the GC content; the internal stability at the 5′ end of the antisense strand; inverted repeats; and base preferences at specific positions, such as positions 3, 10, 13, and 19. They have developed a set of criteria for functional siRNAs. For example, criterion-I stated that 30–52% of GC content in the siRNA sequences is good for functional siRNAs. This criterion alone, however, only marginally improved the selection of functional siRNAs. Criterion-II identified that one or more "A/U" base-pairs in position 15–19 of the sense strand would increase the chance of having functional siRNAs. Again, this criterion alone only marginally increased the selection of functional siRNAs. Criterion-III stated that internal repeat and hairpin forming potentials are not good for functional efficiency. Additional rules were formulated which concern the identity of the NT at specific positions in siRNAs. For example, criterion-IV stated that an "A" at position 19 of the sense strand increased the chance of functional siRNA. Functional siRNAs tend to have an "A" at position 3 of the sense strand and a "U" at position 10 at the sense strand. These were criterion V and VI, respectively. Criterion-VII indicated that the absence of "G" or "C" at position 19 was good; and criterion-VIII stated that absence of "G" at position 13 was beneficial for siRNA's efficacy. A formal scheme was then formulated which combined the above criteria into one coherent scoring system. The weighting in the scoring was somewhat arbitrary, and the scoring system was not for quantitative prediction; rather, it was meant to be used to distinguish potentially functional and nonfunctional siRNAs.

Ui-Tei et al. [17] have looked at siRNAs targeting six genes. Similar to Reynolds, they analyzed the statistical significance of several rules or criteria for functional siRNAs. The rules are: (a) A/U at the 5′ end of the antisense strand; (b) G/C residue at the 5′ end of the sense strand; (c) A/U richness in first seven positions of the antisense 5′ terminal; and (d) the absence of long GC stretch of more than nine NTs in the siRNA sequences are all good for siRNA functionality. Similarly, Amarzguioui et al. [19] described an algorithm to help select functional siRNAs. They have analyzed 46 siRNAs and identified features that correlate

well with efficiency at the 70% knockdown level. They further verified the rules against an independent data set of 34 siRNAs. A selection algorithm was then formulated based on these findings.

Chalk et al. [8] were the first to employ a formal classification tree method with cross-validation to derive rules for the rational design of siRNAs. They developed the models based on a dataset of 398 siRNAs of known efficacy targeting 92 different genes. They have reported that the new design rules could identify functional siRNAs with efficacy greater than 50% in 91% of the cases, and those with efficacy greater than 90% in 52% of the cases.

Other rule based methods were also published based on the analysis of larger datasets. For example, Jagla et al. [20] examined the sequence characteristics of functional siRNAs based on a dataset of 601 siRNAs targeting four different genes. A decision tree algorithm was employed to derive the rules. Relative to the sense strand of the siRNA sequence, the best rules were: (a) an "A/U" at positions 10 and 19; (b) a "G/C" at position 1, and (c) more than three "A/U"s between positions 13 and 19. They claimed that with these rules, there was a 99.9% chance of designing an effective siRNA with more than 50% efficacy.

Although these rules were helpful for understanding and hypothesizing molecular mechanisms involved in the RNAi process, quantitative modeling with these approaches was difficult. The various design guidelines suffer from two problems: they differ considerably from each other, and they produce high levels of false-positive predictions when tested on data from independent sources [21]. More complicated factors are likely involved in determining siRNA's gene silencing potency, and cooperative interactions among the various factors may play an important role in determining the efficacy of siRNAs. Thus, more rational approaches involving machine learning methods and multivariate data analysis may be needed to achieve more robust and predictive results.

2.3.2. Machine Learning Based Modeling

The first machine learning based modeling of siRNA was published by Saetrom [6]. Experimentally determined siRNAs were collected from multiple literature sources. After careful preprocessing, they have obtained a dataset of 204 siRNAs (101 positive and 103 negative examples). They used this dataset to train GPboost, a machine learning algorithm. Several versions of the Support Vector Machine algorithms (SVM) were also employed in their work for comparative analyses. The performance of the developed models was examined in terms of the Area Under the Curve (AUC) derived from the ROC curves as well as Pearson coefficient R between the predicted and the actual efficacy. Several encoding methods for siRNA features were described. The first encoding method simply transformed the sequences into

numerical data by mapping A to 0, C to 0.33, G to 0.67 and T/U to 1. The second encoding uses the relative number of each nucleotide and dinucleotide in the sequence. The third encoding uses the relative number of tetramers in each sequence. The AUCs fell between 0.53 and 0.72; and the Pearson coefficient *R* fell between 0.02 and 0.46, which were not particularly great but represented the first attempt at this challenging topic.

In a follow-up benchmark study, Saetrom and Snove [22] compared different rule-based scoring methods and their newly developed machine learning method (the GPBoost model). A dataset of 581 diverse siRNAs were collected from literature sources, targeting 40 different genes. Included in that benchmark study were GPBoost [6], Ui-Tei's guidelines [17], Amarzguioui [19], Hsieh [23], Takasaki [24], different versions of Reynolds [16], Schwarz [25], Khvorova [26], different versions of the Stockholm algorithms [8], a Tree method [8], and a method published by Luo et al. [27] using mRNA secondary structure to build prediction models. On the basis of three independent test sets, they have concluded that the machine learning method, GPBoost, was the best, which afforded an AUC of 0.84 and a Pearson coefficient *R* of 0.55. The methods by Ui-Tei, Amarzguioui as well as by Reynolds also performed fairly well, while other tested methods fell short of expectation. Because the GPBoost algorithm used only the sequence information of the siRNAs as the input for training and testing, they also concluded that functional siRNA sequences can capture indirectly other factors that may contribute to the efficacy of siRNAs. Thus, it is sufficient to use the siRNA sequence information or its derived information to build predictive models without having to describe explicitly the local structure of the target mRNAs. This is also the fundamental hypothesis of many other algorithms developed since Saetrom's work.

One of the landmark work on siRNA efficacy modeling is published by Huesken et al. [7]. They employed the artificial neural network (ANN) method to develop a predictive model called *BioPredsi*. Subsequently, they employed the model to predict siRNA efficacy of new siRNAs, and validated them by experimental work. In their publication, they trained the model on a data set of 2,182 randomly selected siRNAs targeting 34 mRNAs, assayed through a high-throughput fluorescent reporter gene system. They demonstrated that *BioPredsi* could reliably predict the activity of 249 siRNAs of an independent test set with a Pearson coefficient *R* of 0.66. One of the key contributions from them was the largest dataset made available to the scientific community, which is by far the largest siRNA dataset generated by a single source under consistent conditions. Thus, the dataset has become the standard training set for many new methods.

Shabalina et al. [12] combined several different sets of features, including the thermodynamic features and position-dependent features into their modeling. A dataset of 653 siRNAs were collected from the literature. Eighteen features were identified that correlated well with siRNAs' silencing efficiency. They optimized a neural network model using three parameters characterizing the siRNA sequences. The prediction of the siRNA efficiency for the Huesken dataset achieved a remarkable Pearson coefficient of 0.75.

Another nice modeling work was published by Vert et al. [28] which reported an accurate and interpretable model for siRNA efficacy prediction. Their work was based on the Huesken dataset. They proposed a simple linear model combining features of siRNA sequences. The model performed as well as *BioPredsi* in terms of prediction accuracy. It was also easier to interpret. They quantified the effect of nucleotide preferences at specific positions. They claimed that the sequence motifs alone contain at least as much relevant information for potency prediction as the nucleotide preferences at specific positions. Their best models gave the Pearson coefficient of about 0.67, statistically tied with that of *BioPredsi*.

In a much needed benchmark work by Matveeva et al. [9], the authors collected four independent datasets to validate various methods. The datasets included the Isis set [29], combined Amgen [26]/Dharmacon set [16], Sloan Kettering set [20], and the Huesken set [7]. These datasets contain 67, 238, 601 and 2,431 siRNAs, respectively. They have examined the performance of several published algorithms in terms of the AUC of the ROC curves as well as the prediction Pearson coefficients. The correlation for all methods analyzed was significant, but the best methods were *BioPredsi* [7], *Thermocomposition* [12] and *DSIR* [28], which all had statistically similar performance on all the datasets. They also developed a new method that utilizes linear regression fitting with local duplex stability, nucleotide position-dependent preferences and total G/C content of siRNAs as the input parameters. The new method's discrimination ability of efficient and inefficient siRNAs is comparable with that of the best methods identified, but its parameters are more obviously related to the mechanisms of action in comparison to *BioPredsi*.

Another thread of research papers on siRNA modeling was published by Li's group [21, 30–32]. A key contribution of their work was the integrated database, *siRecords*, collected and annotated from current literature studies. Based on this large dataset (close to 3,000 siRNAs, and the size continues to grow), they have surveyed many reported features associated with high RNAi efficacy. By performing quantitative analyses on cooperative effects among these features, and then applying a disjunctive rule merging (DRM) algorithm, they have developed a bundled rule set

implemented in siDRM. Several filters have also been implemented to check unwanted detrimental effects, including innate immune responses, cell toxic effects and off-target activities in selecting siRNAs. They have reported a comparative analysis of existing algorithms and their method, and concluded that siDRM outperformed 19 other siRNA design tools in identifying effective siRNAs [21]. However, independent benchmark studies may be needed to further substantiate their report.

Lu and Mathews took a different approach to siRNA modeling based on hybridization thermodynamics [11]. The equilibrium of the siRNA-target hybridization was used for selection of efficient siRNAs. They have shown that the accessibilities of siRNA and target mRNA for hybridization are significantly correlated with siRNA efficacy. The predicted thermodynamic features, in addition to siRNA sequence features, were used as input for a support vector machine modeling. The method was reported to work well for selecting efficient siRNAs (>70%) from the Huesken dataset [7]. The positive predictive value was reported to be as high as about 88%.

Several studies employed machine learning methods to build models as well as to characterize the importance of various features for siRNA efficacy. For example, Peek has combined several known important features and conducted feature filtering analysis before model building [33]. Eight overall feature mapping methods were compared in their abilities to build SVM regression models that predict published siRNA activities. The paper concluded that the primary factors in predictive SVM models are (1) position specific nucleotide compositions; (2) sequence motifs and (3) guide strand to passenger strand sequence thermodynamics.

The most recent work on identifying critical features in a systematic fashion was published by Jochen et al. [13]. They attributed the limited success of current siRNA models to small datasets and the heterogeneity of available datasets, and therefore attempted to overcome these problems by constructing a large meta-dataset of 6,483 siRNAs and then applying a Bayesian analysis (which accommodates feature set uncertainty) to derive important features. A stochastic logistic regression-based algorithm was designed to explore 497 compositional, structural and thermodynamic features, in order to identify associations with siRNA potency. This is by far the most comprehensive analysis of different kinds of features that might be related to siRNA efficacy. The features studied by them include position-dependent nucleotide preferences, GC content, presence of 2-, 3- and 4-mer sequence motifs, thermodynamic features, structural features as well as other factors.

To conclude this section, we summarize the main publications in siRNA efficacy modeling in Table 1, where Saetrom's

Table 1
Major work on siRNA efficacy modeling and the timeline of their development

2004	2005	2006	2007	2008	2009
Reynolds' rules	Huesken model	Gong	Matveeva benchmark	siDRM	
Ui-Tei guidelines		Shabalina	Peek model	Lu and Mathews	Klingelhoefer
Amarzguioui		Vert			
Chalk method					
Saetrom model					
Saetrom benchmark					

benchmark, Huesken model and its dataset, Matveeva's benchmark work as well as the siRecords work are the notable milestones.

2.3.3. Models with Significant Mistakes

It is probably important to note that some published models are completely mistaken. Jia et al. [34] published two systems: (1) a statistical model based on sequence information and (2) a machine learning model based on three features of siRNA sequences (binary description, thermodynamic profile and nucleotide composition). They claimed that both methods show high performance with Pearson coefficients suspiciously high (>0.90). A careful analysis of their training set indicated that the Huesken dataset used by these authors was the wrong set first published mistakenly by Huesken et al. (but later corrected). The wrong dataset was not the experimental data, but the predicted efficacy values by the *BioPredsi* algorithm. Thus, Jia's models were constructed to correlate with *BioPredsi*'s predictions rather than real efficacy of the siRNAs, which explains well why their Pearson coefficients were unusually high. Another system published by Jiang et al. [35] based on random forest regression model was similarly wrong based on the wrong dataset. Thus, we caution that if one developed models with predictive Pearson coefficients greater than 0.90 with the Huesken dataset, he/she should double-check whether the correct dataset was being used in training the models.

2.4. Examples of Web-Enabled Tools for siRNA Library Design

In addition to the published algorithms for siRNA modeling, many online siRNA design systems have been implemented.

Yuan et al. described a siRNA prediction program, and it is accessible at the Whitehead Institute server [36]. After entering

the sequence and various filtering methods, the user will be presented with a list of RNAs matching the specified criteria. From the list, one can select RNAs for further consideration. All the selected sequences will be BLASTed against RefSeq, UniGene, or Ensembl databases. Resulted siRNAs can be further filtered to reduce non-specific target effects. *BioPredsi*, as summarized above, can be accessed via the internet. One can upload a particular sequence or an identifier for a given gene, and the software will design a user-specified number of siRNA sequences that are predicted to have an optimal knockdown effect on the target gene. For every siRNA sequence, a score is returned, which reflects the predicted potential of the siRNA to decrease the expression levels of a target mRNA. The *siDESIGN* Center [16] is an another siRNA design tool, which helps improve the chance of identifying functional siRNA. *siDRM* [21] is an implementation of the DRM rule sets for selecting effective siRNAs. These rule sets were obtained in a three step analysis of the siRecord database. Naito et al. [37] described *siVirus*, a web-based tool designed specifically for antiviral siRNA design. It searches for functional, off-target minimized siRNAs targeting highly conserved regions of viral sequences. These siRNAs are expected to resist viral mutational escape, since their highly conserved targets likely contain constrained elements. OligoWalk [38] is another online siRNA design tool utilizing hybridization thermodynamics. Given an mRNA sequence as input, it generates a list of siRNA sequences, ranked by the probability of being efficient siRNA (silencing efficacy greater than 70%). AsiDesigner [39] is an exon-based siRNA design server. It considers alternative splicing. It provides numerous novel functions including the design of common siRNAs for silencing more than two mRNAs simultaneously. All the above tools and their internet addresses are summarized in Table 2.

Table 2
Online tools for rational siRNA design

Method	References	URL
Whitehead	[36]	http://jura.wi.mit.edu/bioc/siRNAext/
BioPredsi	[7]	http://www.biopredsi.org/start.html
siDESIGN	[16]	http://www.dharmacon.com/designcenter/designcenterpage.aspx
siDRM	[21]	http://sirecords.umn.edu/siDRM/
siVirus	[37]	http://sivirus.rnai.jp/
OligoWalk	[38]	http://rna.urmc.rochester.edu/servers/oligowalk
AsiDesigner	[39]	http://sysbio.kribb.re.kr:8080/AsiDesigner/menuDesigner.jsf.

3. Cheminformatics Approach to siRNA Modeling

Rational design of siRNAs can be aided by multivariate machine learning models as demonstrated by many publications cited above, without a full understanding of why each factor is important for siRNA's efficacy. This is not to underestimate the importance of understanding the molecular mechanisms involved in RNAi. But it does indicate that statistical analysis can be employed to capture the correlative relationship between sequence features and the efficacy of siRNAs. State-of-the-art models or rules developed for siRNA efficacy prediction all employ sequence information (such as position-specific NT preferences and sequence motifs) as well as sequence-derived properties (such as thermodynamic quantities) as the basis for modeling. No methods based on cheminformatics or QSAR methods have been published thus far. Since a chemical approach has the potential to describe the structural relationships between nucleotide residues, properly designed chemical descriptors may be useful for quantitative modeling of siRNA efficacy. Furthermore, current models are all based on siRNAs consisting of natural nucleotide residues. Therefore, they are not able to predict or model chemically modified siRNAs, which has a great potential to improve the pharmacokinetic properties of siRNAs used as therapeutic agents. On the other hand, cheminformatics methods can describe siRNAs, natural or chemically modified, in a consistent fashion. Thus, it is of interest to apply chemical descriptors in siRNA modeling.

Here, we briefly outline a new approach to siRNA modeling using cheminformatics descriptors. We have adopted descriptors similar to those for peptide modeling [40] or modeling of DNA molecules [41]. We first calculate 327 common chemical descriptors in MOE (Molecular Operating Environment) for each of the four nucleotide molecules. A principal component analysis (PCA) of the 4 by 327 matrix has led to three principal components explaining 100% of the variance. As a result, Table 3 was generated, where each nucleotide residue is described by three numbers (the Z descriptors). Thus, a given siRNA sequence can be described by a set of numerical descriptors by simply substituting each NT by its three numerical Z descriptors. For example, a 19 NT sequence would be converted to 57 numerical descriptors, and a 21 NT siRNA sequence would be converted to 63 descriptors. The descriptor generation workflow is summarized in Fig. 3.

The Huesken dataset was used in this report. The dataset was split according to Huesken et al. into training and test sets. The 19 NT representation of the siRNA sequences without the 2 NT overhang was used. As explained, each siRNA sequence has been converted into 57 numerical descriptors. Once the dataset has

Table 3
Principal component scores of naturally occurring nucleotides in RNA

Nucleotide	*z*1	*z*2	*z*3
Adenine (A)	0.809	1.47	0.430
Uracil (U)	−1.156	−0.403	1.225
Guanine (G)	1.161	−1.278	−0.141
Cytosine (C)	−0.813	0.211	−1.515

The PCA scores are obtained by performing PCA analysis of the data matrix formed by the four nucleotides (A, U, G, and C) and their 327 calculated molecular descriptors with the MOE program

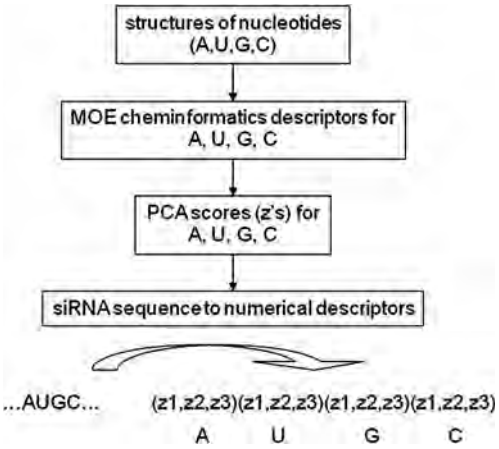


Fig. 3. Cheminformatics descriptors for a siRNA molecule.

been converted to the numerical descriptors, the problem has become a standard QSAR modeling problem, where the activity is the gene silencing potency of the siRNAs. Support Vector Machine (SVM) regression method has been used to train the QSAR models. Figure 4 shows the scatter plots for four test cases, where the QSAR model predicted versus actual siRNA potency on the test set are displayed. The Pearson coefficients for the test set range from 0.59 to 0.64. Thus, the best model affords similar prediction accuracy as one of the best models, *BioPredsi*. Further optimization of the method has led to better models that will be described in more details elsewhere.

Although this method only affords models with similar prediction accuracy, it demonstrates that chemical descriptors can be used in building predictive models for siRNA efficacy. The potential advantage of this approach is that it lays a foundation for modeling

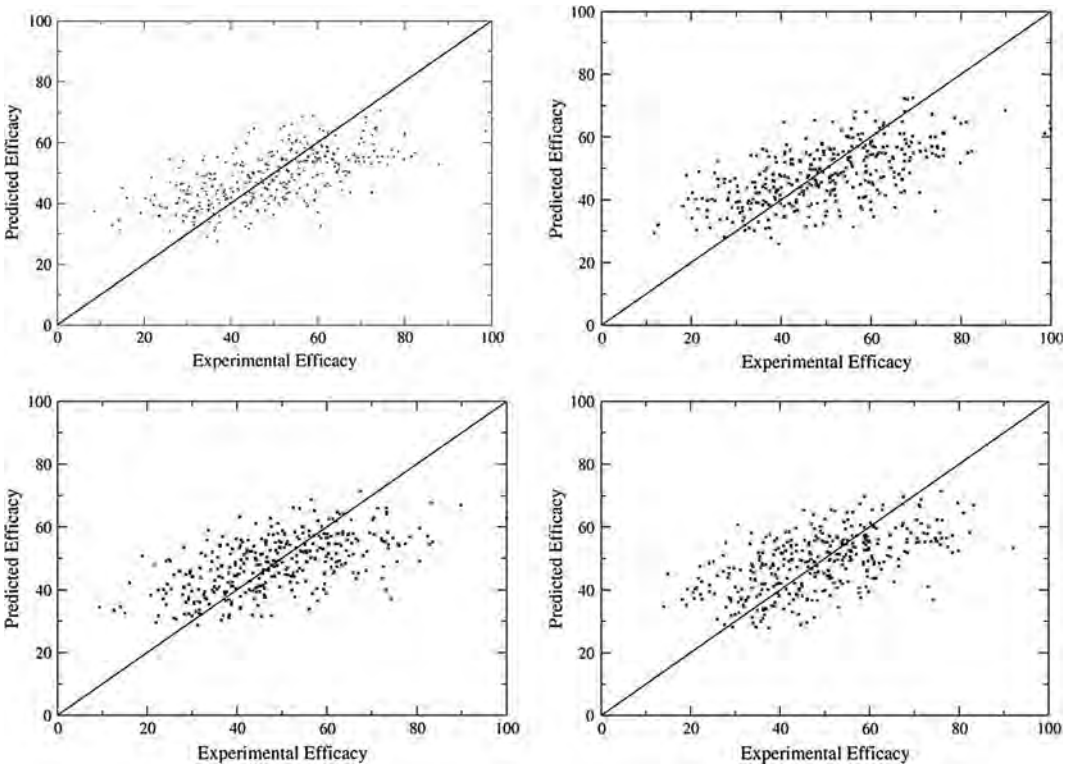


Fig. 4. Scatter plots for model-predicted potency against experimental potency for four of the test sets. The *diagonal lines* are drawn to indicate where the perfect correlation lines would be. Pearson correlation coefficients for the four plots are 0.59, 0.64, 0.64, and 0.62. These results were obtained using the 19 NT representation of the siRNAs.

chemically modified siRNAs. Our preliminary results indicate that high correlation can be obtained with this approach (data not shown) to model chemically modified siRNAs. Thus, we expect to see more development of cheminformatics approaches to the modeling of siRNA efficacy and contribute to the rational design of chemically modified siRNAs as potential RNAi therapeutics.

4. Concluding Remarks

siRNA technology represents a powerful approach to modulate posttranscriptional gene expression. It has been demonstrated as a critical technology for target identification and target validation in the drug discovery industry as well as pathway elucidation in systems biology research. To effectively utilize this powerful technology, rational methods for siRNA design are highly desirable.

We have summarized major modeling work on siRNA design and highlighted several milestones in the development of predictive models for siRNA efficacy. These examples prove that siRNA

sequences contain sufficient information for predictive modeling of its efficacy. As demonstrated in several benchmark studies, existing tools are effective in predicting and selecting highly efficient siRNAs. Online tools are becoming accessible to the scientific community. As more RNAi data becomes available, further refinement of the rules or models would afford more predictive models for siRNA library design.

As chemically modified siRNAs tend to improve their pharmacokinetic properties, we begin to witness more RNAi experiments performed with these modified siRNAs and their potential to be used as RNAi based therapeutics. For this reason, it is desirable to develop new cheminformatics approaches to siRNA modeling in order to complement the traditional bioinformatics approaches which are based solely on sequences of natural nucleotides. Our new cheminformatics method has already shown its promise in modeling chemically modified siRNAs. Together with novel delivery technologies for siRNAs, these new algorithms are expected to contribute to the rational development of siRNAs into potential therapeutic agents.

Acknowledgment

We acknowledge the financial support of the Golden Leaf Foundation via the BRITE Institute, North Carolina Central University.

References

1. Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811.
2. Mello, C. C., and Conte, D., Jr. (2004) Revealing the world of RNA interference. *Nature* **431**, 338–342.
3. Gao, K., and Huang, L. (2009) Nonviral methods for siRNA delivery. *Molecular Pharmaceutics* **6**, 651–658.
4. Bramsen, J. B., Laursen, M. B., Nielsen, A. F., Hansen, T. B., Bus, C., Langkjaer, N., Babu, B. R., Hojland, T., Abramov, M., Van Aerschot, A., Odadzic, D., Smicius, R., Haas, J., Andree, C., Barman, J., Wenska, M., Srivastava, P., Zhou, C., Honcharenko, D., Hess, S., Muller, E., Bobkov, G. V., Mikhailov, S. N., Fava, E., Meyer, T. F., Chattopadhyaya, J., Zerial, M., Engels, J. W., Herdewijn, P., Wengel, J., and Kjems, J. (2009) A large-scale chemical modification screen identifies design rules to generate siRNAs with high activity, high stability and low toxicity. *Nucleic Acids Research* **37**, 2867–881.
5. Patzel, V. (2007) In silico selection of active siRNA. *Drug Discovery Today* **12**, 139–148.
6. Saetrom, P. (2004) Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics* **20**, 3055–3063.
7. Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D., Labow, M., Reinhardt, M., Natt, F., and Hall, J. (2005) Design of a genome-wide siRNA library using an artificial neural network. *Nature Biotechnology* **23**, 995–1001.
8. Chalk, A. M., Wahlestedt, C., and Sonnhammer, E. L. (2004) Improved and automated prediction of effective siRNA. *Biochemical and Biophysical Research Communications* **319**, 264–274.

9. Matveeva, O., Nechipurenko, Y., Rossi, L., Moore, B., Saetrom, P., Ogurtsov, A. Y., Atkins, J. F., and Shabalina, S. A. (2007) Comparison of approaches for rational siRNA design leading to a new efficient and transparent method. *Nucleic Acids Research* **35**, e63.
10. Chan, C. Y., Carmack, C. S., Long, D. D., Maliyekkel, A., Shao, Y., Roninson, I. B., and Ding, Y. (2009) A structural interpretation of the effect of GC-content on efficiency of RNA interference. *BMC Bioinformatics* **10(Suppl 1)**, S33.
11. Lu, Z. J., and Mathews, D. H. (2008) Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Research* **36**, 640–647.
12. Shabalina, S. A., Spiridonov, A. N., and Ogurtsov, A. Y. (2006) Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics* **7**, 65.
13. Klingelhoefer, J. W., Moutsianas, L., and Holmes, C. (2009) Approximate Bayesian feature selection on a large meta-dataset offers novel insights on factors that effect siRNA potency. *Bioinformatics* **25**, 1594–1601.
14. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. (2001) Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *The EMBO Journal* **20**, 6877–6888.
15. Elbashir, S. M., Harborth, J., Weber, K., and Tuschl, T. (2002) Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods* **26**, 199–213.
16. Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S., and Khvorova, A. (2004) Rational siRNA design for RNA interference. *Nature Biotechnology* **22**, 326–330.
17. Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., Ueda, R., and Saigo, K. (2004) Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Research* **32**, 936–948.
18. Holen, T. (2006) Efficient prediction of siRNAs with siRNArules 1.0: an open-source JAVA approach to siRNA algorithms. *RNA* **12**, 1620–1625.
19. Amarzguioui, M., and Prydz, H. (2004) An algorithm for selection of functional siRNA sequences. *Biochemical and Biophysical Research Communications* **316**, 1050–1058.
20. Jagla, B., Aulner, N., Kelly, P. D., Song, D., Volchuk, A., Zatorski, A., Shum, D., Mayer, T., De Angelis, D. A., Ouerfelli, O., Rutishauser, U., and Rothman, J. E. (2005) Sequence characteristics of functional siRNAs. *RNA* **11**, 864–872.
21. Gong, W., Ren, Y., Zhou, H., Wang, Y., Kang, S., and Li, T. (2008) siDRM: an effective and generally applicable online siRNA design tool. *Bioinformatics* **24**, 2405–2406.
22. Saetrom, P., and Snove, O., Jr. (2004) A comparison of siRNA efficacy predictors. *Biochemical and Biophysical Research Communications* **321**, 247–253.
23. Hsieh, A. C., Bo, R., Manola, J., Vazquez, F., Bare, O., Khvorova, A., Scaringe, S., and Sellers, W. R. (2004) A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Research* **32**, 893–901.
24. Takasaki, S., Kotani, S., and Konagaya, A. (2004) An effective method for selecting siRNA target sequences in mammalian cells. *Cell Cycle* **3**, 790–795.
25. Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P. D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208.
26. Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216.
27. Luo, K. Q., and Chang, D. C. (2004) The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochemical and Biophysical Research Communications* **318**, 303–310.
28. Vert, J. P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. (2006) An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics* **7**, 520.
29. Vickers, T. A., Koo, S., Bennett, C. F., Crooke, S. T., Dean, N. M., and Baker, B. F. (2003) Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *The Journal of Biological Chemistry* **278**, 7108–7118.
30. Gong, W., Ren, Y., Xu, Q., Wang, Y., Lin, D., Zhou, H., and Li, T. (2006) Integrated siRNA design based on surveying of features associated with high RNAi effectiveness. *BMC Bioinformatics* **7**, 516.
31. Ren, Y., Gong, W., Xu, Q., Zheng, X., Lin, D., Wang, Y., and Li, T. (2006) siRecords: an extensive database of mammalian siRNAs with efficacy ratings. *Bioinformatics* **22**, 1027–1028.
32. Ren, Y., Gong, W., Zhou, H., Wang, Y., Xiao, F., and Li, T. (2009) siRecords: a database of mammalian RNAi experiments and efficacies. *Nucleic Acids Research* **37**, D146–D149.

33. Peek, A. S. (2007) Improving model predictions for RNA interference activities that use support vector machine regression by combining and filtering features. *BMC Bioinformatics* **8**, 182.
34. Jia, P., Shi, T., Cai, Y., and Li, Y. (2006) Demonstration of two novel methods for predicting functional siRNA efficiency. *BMC Bioinformatics* **7**, 271.
35. Jiang, P., Wu, H., Da, Y., Sang, F., Wei, J., Sun, X., and Lu, Z. (2007) RFRCDB-siRNA: improved design of siRNAs by random forest regression model coupled with database searching. *Computer Methods and Programs in Biomedicine* **87**, 230–238.
36. Yuan, B., Latek, R., Hossbach, M., Tuschl, T., and Lewitter, F. (2004) siRNA selection server: an automated siRNA oligonucleotide prediction server. *Nucleic Acids Research* **32**, W130–W134.
37. Naito, Y., Ui-Tei, K., Nishikawa, T., Takebe, Y., and Saigo, K. (2006) siVirus: web-based antiviral siRNA design software for highly divergent viral sequences. *Nucleic Acids Research* **34**, W448–W450.
38. Lu, Z. J., and Mathews, D. H. (2008) OligoWalk: an online siRNA design tool utilizing hybridization thermodynamics. *Nucleic Acids Research 36*, W104–W108.
39. Park, Y. K., Park, S. M., Choi, Y. C., Lee, D., Won, M., and Kim, Y. J. (2008) AsiDesigner: exon-based siRNA design server considering alternative splicing. *Nucleic Acids Research* **36**, W97–W103.
40. Sandberg, M., Eriksson, L., Jonsson, J., Sjostrom, M., and Wold, S. (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *Journal of Medicinal Chemistry* **41**, 2481–2491.
41. Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., and Wold, S. (1993) Quantitative sequence-activity models (QSAM) – tools for sequence design. *Nucleic Acids Research* **21**, 733–739.

Chapter 15

Beyond Rhodopsin: G Protein-Coupled Receptor Structure and Modeling Incorporating the β2-adrenergic and Adenosine A_{2A} Crystal Structures

Andrew J. Tebben and Dora M. Schnur

Abstract

For quite some time, the majority of GPCR models have been based on a single template structure: dark-adapted bovine rhodopsin. The recent solution of β2AR, β1AR and adenosine A_{2A} receptor crystal structures has dramatically expanded the GPCR structural landscape and provided many new insights into receptor conformation and ligand binding. They will serve as templates for the next generation of GPCR models, but also allow direct validation of previous models and computational techniques. This review summarizes key findings from the new structures, comparison of existing models to these structures and highlights new models constructed from these templates.

Key words: G-Protein coupled receptor, GPCR, Homology model, Crystal structure, Adrenergic receptor, Adenosine receptor

1. Introduction

G-protein coupled receptors (GPCRs) are a family of 7-transmembrane integral membrane proteins with an extracellular N-terminus, three extracellular loops, three intracellular loops, and an intracellular C-terminus (Fig. 1a, b). Their primary function is the transmission of extracellular signals via coupling to intracellular trimeric G-proteins, which trigger a variety of responses including calcium release, cyclic-AMP synthesis, and activation of phospholipase Cβ, amongst others. They are the largest gene super family within the human genome, comprising 720–800 genes or ~2% of the genome [1]. This broad array of sequences potentiates cellular responses to a diverse set of stimuli ranging from small molecules, peptides, fatty acids to proteins. Since GPCRs have been shown to play a central role in multiple disease processes including inflammation [2], viral attachment [3], diabetes [4], and hypertension [5], this protein family contains numerous desirable targets for therapeutic intervention. Drug

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_15, © Springer Science+Business Media, LLC 2011

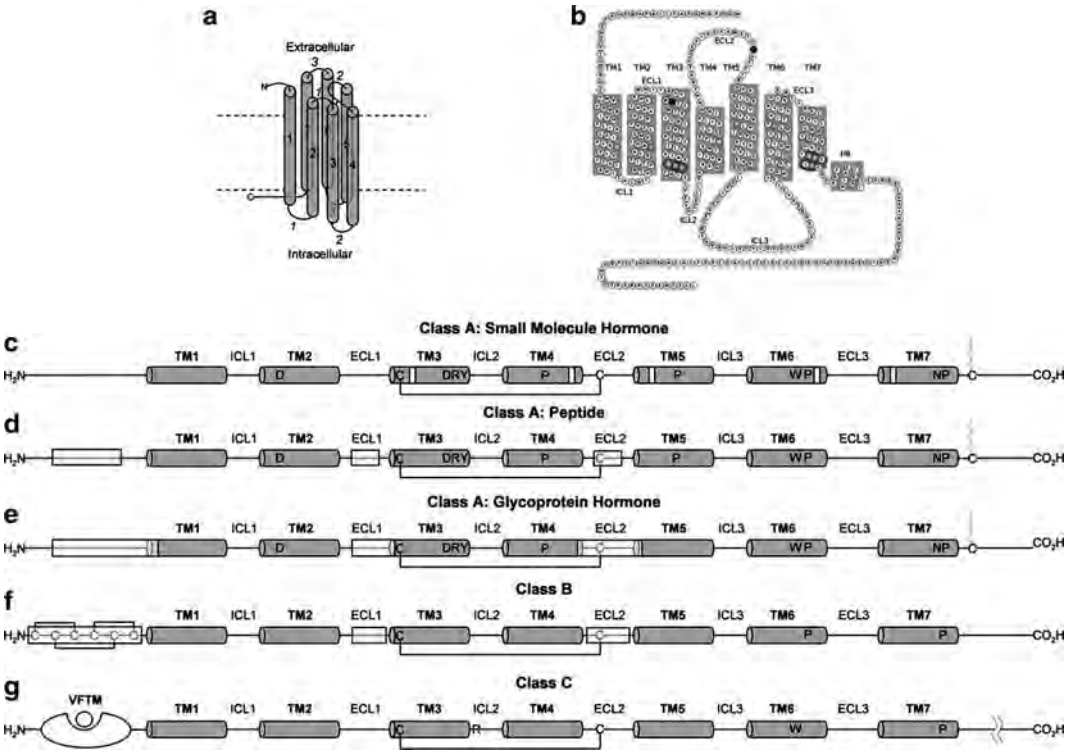


Fig. 1. (**a**) Schematic representation of GPCR topology. Helices are numbered within the cylinders. Intra- and extracellular loop numbers are in italics. *Dashed lines* represent the membrane. (**b**) An example of the commonly used GPCR snake plot generated by unfolding the receptor bundle and flattening it in two dimensions. In the β2AR plot shown, the most conserved residues within the Class A family are denoted with *gray circles*, the cystienes forming the disulfide linkage between TM3 and ECL2 are shown in *black*. (**c**–**g**) Linear arrangement of GPCR structural elements. Conserved residues for each class and subclass are denoted by single letter residue codes within each diagram. Areas involved in ligand binding are indicated with *white* boxes. The ligand binding region for Class A, small molecule hormone receptors (**c**) is within the trans-membrane bundle. Class A peptide and glycoprotein hormone binding elements are located on the N-terminal domains, extracellular loops and the extracellular ends of TM1, 3, 4, and 5 (**d**–**e**). Class B receptors have a structured N-terminal domains containing three disulfide bonds. Ligand binding is primarily within this domain with additional contributions from ECL1 and ECL2 (**f**). The large N-terminal domain of the Class C receptors carries the Venus flytrap module which binds small molecule agonists (**g**).

discovery efforts in this area have been quite productive as GPCR targeted drugs now comprise ~30% of all marketed drugs [6].

The GPCR family has been broadly classified into three subfamilies based on primary sequence [7], with members within each subfamily sharing over 25% sequence homology within the trans-membrane region. Class A, the rhodopsin like family, is the largest and best understood in terms of structure and function. This family is characterized by several highly conserved residues within the trans-membrane helical regions, a palmitoylated cysteine near the C-terminus, and usually a disulfide link between extracellular loop 2 (ECL2) and the top of trans-membrane helix 3 (TM3). The small molecule ligand biding site is generally found within the helical bundle toward the extracellular side,

while those that bind peptides and glycoprotein hormones derive significant binding affinity from interactions with the N-terminal domain and the extracellular loops (Fig. 1 – Class A). However, even within the peptide binding receptors, it has been shown that the actual signaling event requires binding within the trans-membrane region. For example, removing residues 2–8 from the N-terminus amino of the endogenous CCR2 agonist MCP-1 converts it into an antagonist [8]. Mutagenesis studies have localized the residues contacting the MCP-1 N-terminus to the helices [9], suggesting that the agonist "address" recognizes the receptor extracellular loops with the "message" binding within the trans-membrane region. Small molecule ligands have been developed for the chemokine family and other peptide receptors that have been shown via mutagenesis to contact helical residues, demonstrating the conservation of a binding site within the trans-membrane domain across the family. Class B receptors lack the characteristic conserved residues from Class A as well as the C-terminal palmityolation site (Fig. 1 – Class B). Instead, they are typified by a large, globular N-terminal domain which serves as the binding site for peptide hormones such as secretin, calcitonin, CRF and GLP [10]. Trans-membrane segments 6 and 7 also bear highly conserved proline residues, likely required for conformational changes on activation. As with the peptide subfamily within Class A, it appears that the N-terminal domain drives ligand binding affinity, but receptor activation requires some contact within the trans-membrane region. For example, N-terminal truncations of CRF [11] and GLP [12] retain potent affinity against their cognate receptors, but they revert from agonists to antagonists. The Class C receptors have large N- and C-terminal domains and are found as dimers. The N-terminal domain has the Venus flytrap module (VTFM) which carries the ligand binding site (Fig. 1 – Class C). Although these receptors are activated by small molecules like γ-amino butyric acid (GABA), glutamate, and calcium, their mechanism of activation is quite different than for Class A. Crystallographic studies of the $mGluR_1$ N-terminal domain have shown considerable reorganization of the VTFM on ligand binding [13]. Current models of Class C receptor signaling predict that the VTFM conformational change results in a global change to the organization of the receptor dimer leading to activation [14].

Given the significance of the GPCR family of proteins, it is not surprising that they have been well studied and a significant body of knowledge exists about their pharmacology and biological function. However, because they are integral membrane proteins with substantial conformational heterogeneity and relatively low levels of expression it has been extremely challenging to elucidate structural information at the atomic level. Data such as this has proven to be extremely valuable in the context of drug discovery

as a three dimensional understanding of the binding site greatly enhances one's ability to rationally design ligands. For quite some time GPCR structure based drug design was based largely on models derived from rhodopsin: initially the low resolution cryoelectron microscopy based structure [15] and later the groundbreaking dark adapted bovine rhodopsin crystal structure [16]. A number of models generated through comparative modeling starting from the rhodopsin or bacteriorhodopsin templates have been used to rationalize mutagenesis and SAR data with varying degrees of success [reviewed in [17] and [18]]. More recently, structures of several members of the class A family, β2-adrenergic [19, 20], β1-adrenergic [21], and adenosine-A_{2A} [22], have been solved. These structures provide more relevant templates for modeling within class A and additional insight into ligand binding showing surprising diversity in their placement within the receptor (Fig. 2a). Comparison of the existing models to these new structures also presents the opportunity to validate the modeling procedures and improve them. As GPCR modeling, particularly in

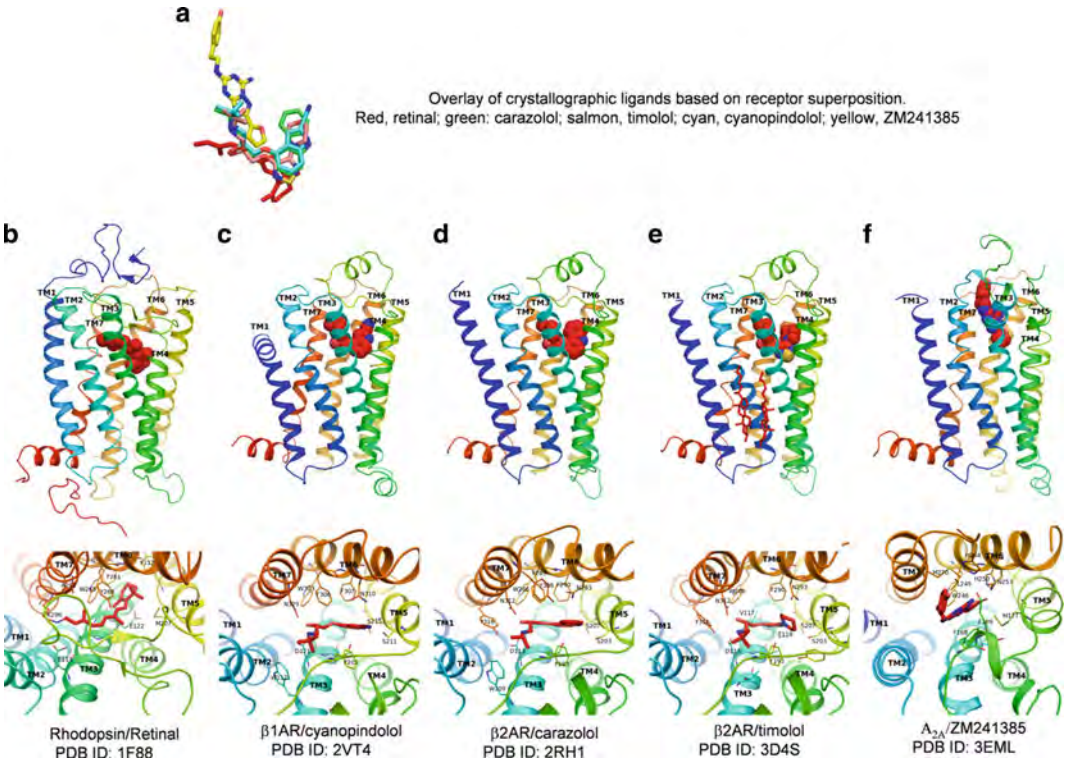


Fig. 2. (**a**) Overlay of crystallographic ligands. Receptors were structurally aligned and the ligand positions extracted from the resulting superposition. (**b**–**f**) Receptor structures and binding sites for bovine rhodopsin/retinal, 1F88 (**b**) turkey β1AR/cyanopindolol, 2VT4 (**c**) β2AR/carazolol, 2RH1 (**d**) β2AR/timolol, 3D4S (**e**) adenosine A_{2A}/ZM241385, 3EML (**f**). Receptors are presented in the same orientation. Helical coloring is a gradient based on residue position from N-terminal, *blue*, to C-terminal, *red*. Bound ligands are shown as spheres except for the cholesterol molecules in E, which are shown in sticks. The T4L insertion is not shown in *panels* **d**, **e**, and **f.**

the context of rhodopsin , has been the subject of several excellent reviews [see [17] and [18]], this review will focus on the recent modeling work derived from the β2 and A_{2A} structures with reference to rhodopsin based modeling only as it relates to these new templates.

2. Structure and Modeling in the Rhodopsin Era

The earliest structural view of a heptahelical receptor was the electron cryo-microscopy based model of bacteriorhodopsin [23]. Although this was a low resolution structure with no direct sequence homology to eukaryotic GPCRs, it did make available the overall topology of the receptor. This data was used as the basis for some very early modeling studies, including angiotensin II [24] and the cationic amine receptors 5-HT2, dopaminergic D2, muscarinic M2 [25]. Low [26] and high [27] resolution crystal structures of bateriorhodopsin later provided more atomic detail, but the challenge of alignment remained. The comparison of bacteriorhodopsin to the subsequent structures of rhodopsin [28], it made it clear that the helices are positioned quite differently, bringing into question the accuracy of bacteriorhodospin derived GPCR homology models. Schertler's [15] subsequent projection structure of rhodopsin and Baldwin's [29] associated model provided the first low resolution view of a GPCR. As a true GPCR, this model was a more relevant template than bacteriorhodopsin, but the level of detail in models built from it was limited due to the lack of atomic detail in the projection data.

The first high resolution GPCR template became available with the solution of the 2.8Å structure of bovine rhodopsin in a dark adapted (inactive) state [16]. The structure not only clearly resolved the trans-membrane domains, but also the intracellular and extracellular loops, and gave an atomic level view of the ligand binding site. Rhodopsin has subsequently been revisited by multiple labs resulting in structures with alternate crystal forms [30], higher resolution [31, 32], and additional species [33]. The recent work has clarified the role of waters and added additional resolution within the loops, but the conformation of the trans-membrane region and retinal binding site are unchanged. Retinal occupies a binding site located about one third of way into the membrane bundle relative to the extracellular side of the receptor (Fig. 2b). It is covalently linked to Lys296 from TM7 forming a Schiff base whose charge is counterbalanced by Glu113 from TM3. The β-ionone ring lies in a hydrophobic pocket contacting residues Trp265 and Tyr268 in TM6 and Phe212 in TM5. Of these, the most significant interaction is Trp265 as the conformation of this residue acts as a switch between the activated and inactivated form

of the receptor. The remainder of the binding pocket is formed by residues from TM3 to TM7. The binding site is capped by a plug formed by extracellular loop 2 (ECL2), which is stabilized by the formation of an extended beta sheet with the N-terminus. The trans-membrane bundle and the ECL2 plug pack tightly around retinal resulting in a tight retinal shaped binding site.

Rhodopsin has very low sequence homology, on the order of 20%, to the non-visual class A receptors of interest, but alignment is possible due to the presence of conserved anchor residues in each trans-membrane region (Fig. 3). Locking these residues in the alignment and using the structure to determine the helical boundaries allows one to align the helical regions with some degree of certainty. Other than the conserved disulfide link between TM3 and EC2, no such anchor residues exist in the loop regions which makes their alignment much more ambiguous. Because of this, many of the rhodopsin based models only detail the trans-membrane regions and exclude the extracellular loops. Several critical interactions are present in the trans-membrane

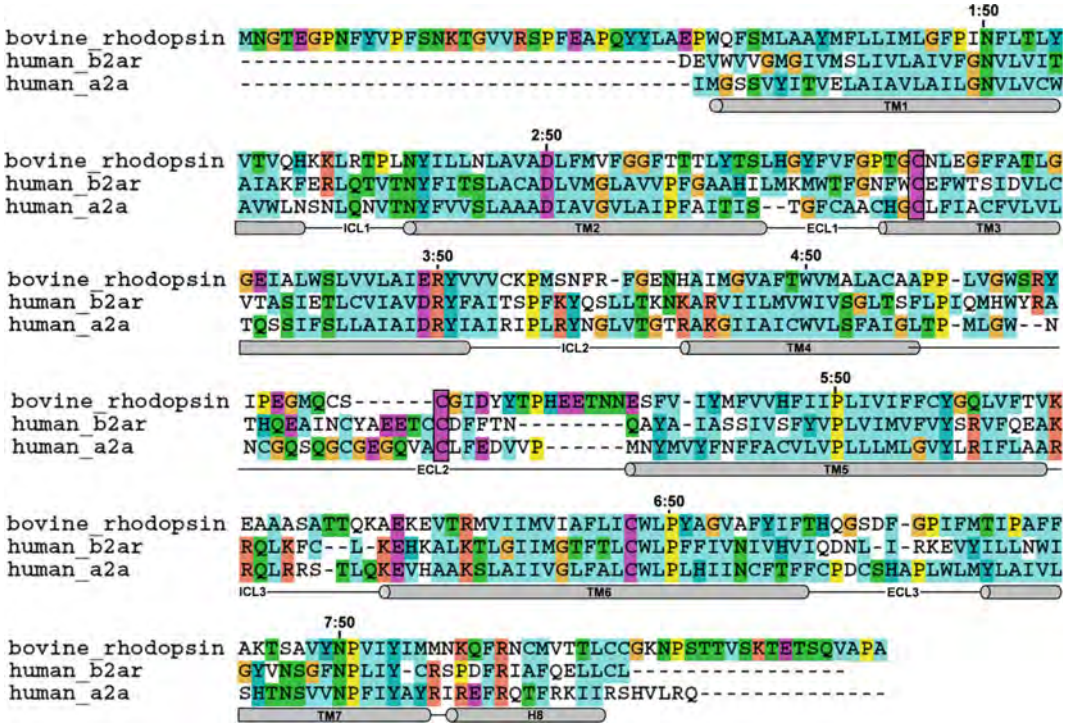


Fig. 3. Structure based alignment of bovine rhodopsin, β2AR, and adenosine A_{2A} sequences from 1F88, 2RH1 and 3EML, respectively. For clarity, the T4 insertions in 2RH1 and 3EML are not shown in the alignment. Cylinders shown below the sequences indicate trans-membrane regions and solid lines the loops. Residues are colored by type: *cyan* – hydrophobic/aromatic, *green* – neutral polar, *purple* – acidic, *orange* – basic, *yellow* – proline, *light orange* – glycine. The most conserved residue in each helix and its corresponding Ballesteros–Weinstein loci are indicated above the sequences. The conserved cysteines that connect TM3 and ECL2 are boxed. These residues serve as anchor points for sequence alignments of Class A family members.

region, but as is demonstrated by the rhodopsin structure itself, ECL2 can play a significant role in defining the boundaries of the binding site as well as possessing ligand-receptor interactions.

Although the topology of the rhodopsin structure is likely representative of the class A family, it represents only a single conformation of a flexible protein. It is of a dark-adapted, inactive state with the inverse agonist retinal bound. This suggests that it may be relevant to antagonist modeling, but brings into question its utility for agonist modeling. Comparison of the light activated rhodopsin structure with the dark adapted revealed that the ligand binding site is largely unperturbed [34] suggesting that if an understanding of ligand binding is the primary goal of the modeling study, rhodopsin may still be a relevant template for both agonists and antagonists. A larger issue is the tight packing of the receptor around its cognate ligand. Models generated from this template will have a similarly small binding cavity which precludes docking of many ligands due to steric issues. In order to generate an accessible binding site, the receptor conformation must be often be altered, either in a systematic way [35] or though manual intervention guided by experimental data [36]. The conformation of extracellular loop 2 is particularly problematic as it protrudes into the trans-membrane region, thereby blocking part of the binding site. Nevertheless, even with these issues, multiple rhodopsin based models have been constructed and validated with available biophysical data. For example, the comprehensive study reported by Bissantz et al. [37], contained models of dopamine D3, muscarinic M1, vasopressin V1a, β2-adrenergic, and δ-opioid receptors. These were used to screen six 1,000-compound 3D databases, each containing 990 random compounds and ten known actives for each target. The screens were able to identify known antagonists, but failed to retrieve known agonists. A rhodopsin based model of GPR40, a member of the fatty acid receptor family was used to identify 15 active compounds whose binding modes were subsequently validated by mutagenesis studies [38]. The chemokine family has also been extensively studied, with antagonist models of CCR1, CCR2, CCR5, and CXCR5 present in the literature [39]. .Several additional examples are cited in recent reviews [see [17, 18]]. Notably, while models have been created that are consistent with the data used to generate them, there are few reports of these models being used to predicatively design compounds.

The conserved residues within the Class A family coupled are the basis for a system of receptor independent residue numbering, referred to as Ballesteros–Weinstein numbering [40]. The most conserved residue in each helix is arbitrarily set to 50 (see Fig. 3), residues N-terminal to this position are decremented by their distance from residue 50 and positions C-terminal are incremented. Residues are then identified by helix number followed by the

residue position. In this text, amino acids will be denoted by their identity and actual sequence number in normal text with the Ballesteros–Weinstein numbers in superscript.

3. Beyond Rhodopsin: The β2, β1, and A_{2A} Structures

The rhodopsin structure made GPCR modeling accessible and for eight years remained the only template available for comparative modeling. Recently, however, new chapters have opened with the crystal structures of the β2-adrenergic, β1-adrenergic, and adenosine A_{2A} receptor. They have revealed details about ligand binding, extracellular loop conformation and helical positioning that could not have been derived from rhodopsin. These structures not only provide a new set of templates but also direct experimental data for the validation of modeling and docking studies.

There are several challenges in producing crystals of GPCRs. These include low levels of expression, lack of crystal contacts and conformational heterogeneity [41]. Rhodopsin is unique in that it is highly expressed and, critical to its function, exists in a stable inactive conformation. In order for the visual system to achieve sufficient sensitivity, rhodopsin must have very low levels of basal signaling in the absence of activation by a photon. This is indeed the case, as the inactive form of the receptor shows no activation of transducin [42]. In contrast, many other GPCRs show basal activity independent of agonist activation which suggests some inherent structural flexibility [43]. In order to crystallize these proteins, the receptor conformation must be stabilized. Two approaches have proven successful; identification of thermally stable mutant receptors via a comprehensive mutational scan of the trans-membrane segments and reducing conformational heterogeneity through truncation of unstructured regions and manipulation of intracellular loop 2. To date, the first strategy has only been applied to the turkey β1-adrengic receptor, while the second has produced structures of the β2-adrenergic and adeonsine A_{2A} receptors.

The structural efforts toward the turkey β1-adrengic receptor began with the postulate that improving its detergent stability would lead to a more crystallizable form [44]. The search for stability took the form of an alanine scan of all residues within the trans-membrane region and 23 from the C-terminus. The resulting mutant receptors were screened for an increase in thermal stability based on their ability to retain >50% of native $\beta1AR_{34\text{-}424}$ antagonist binding affinity after being heated to 32°C for 30 min (the apparent T_m for the native receptor). Eighteen mutants showed an increase in stability and expressed well. These 18 positions were then mutated to 2–5 additional

amino acids, five were improved, 12 showed no additional improvement, and 1 was a false positive. Further optimization was achieved through combination of the individual mutations ultimately leading to a receptor with an apparent T_m of 52.8°C. Although antagonist binding to the mutant receptor was similar to wild-type, agonist binding was significantly altered, with a 2,470- and 650-fold loss in norepinephrine and isoprenaline binding, respectively; indicating that the receptor is stabilized in an antagonist bound conformation. A 2.4-Å crystal structure of this receptor in complex with the antagonist cyanopindolol was obtained [21] (Fig. 2c) with receptor conformation and binding site interactions that were very similar to that of β2AR, which had been released immediately before it (see below).

The Kobilka lab took a more systematic approach toward stabilizing the β2-adrenergic receptor. The structure of a β2AR/Fab in complex with carazolol, reported by Rasmussen et al. [19], was the product of a long investigation within the Kobilka lab. It had been known for quite some time that it was possible to express β2AR at high levels and to purify functional receptors [45]. But, several studies had shown that the β2AR receptor has significant basal activity and undergoes conformational changes on agonist and antagonist binding [46], which likely hindered early crystallization attempts. To stabilize the receptor conformation, several areas of potential heterogeneity were identified: post-translational modification, unstructured sequences in the C-terminus and the large intracellular loop 3 (ICL3). Rasmussen and his co-workers addressed the first two areas by enzymatically removing glycosylation sites and truncating the C-terminus. Association with a Fab stabilized the third intracellular loop and provided the polar contact surface required for crystal contacts and a stable crystal lattice. The receptor conformation was also stabilized by the presence of the inverse agonist carazolol. This combination of factors resulted in a structure with 3.5 Å resolution in the plane of the membrane and 3.7 Å perpendicular to it. As expected, the structure revealed that the crystal contacts were within the Fab region, with no contacts between the receptors themselves. Unfortunately, while the Fab and the intracellular side of the receptor are well resolved, the density degrades toward the extracellular face with only sparse density around the ligand and leaving the extracellular loops unresolved. The structure was sufficiently well resolved to show that the carazolol and retinal binding sites are co-localized and that the overall receptor topology is similar to rhodopsin with a Cα RMSD of 1.56 Å in the trans-membrane region.

The theme of stabilizing receptor conformation through ICL3 was continued with the subsequent report of a β2/carazolol complex at 2.4 Å [20, 41]. In this work, ICL3 is truncated and replaced with a more cystallizable entity, T4 lysozyme (T4L). The key concern about such a radical modification is its potential effect

on the pharmacology of the receptor because introducing a relatively rigid globular protein in place of a flexible loop might alter the conformational manifold of the receptor. Along the same lines, there is the possibility that the T4L might lock out some conformations, resulting in an inaccurate picture of the receptor structure. Competition binding studies with the inverse agonist ICI-118,551 showed no change in its affinity to the T4L construct when compared to wild type. The affinities for the agonist (−)-isopreternol and partial agonist salbutamol were two-to three fold higher for β2AR-T4L versus wild type. The ability of β2AR-T4L to change conformation on ligand binding was probed by attaching a fluorophore to Cys265 at the intracellular end of helix VI, which had previously been shown to detect conformational changes on agonist binding. A qualitatively similar decrease in fluorescent intensity and a shift in the wavelength at maximum intensity (λ_{max}) were observed on isopreternol and salbutamol binding to β2AR-T4L when compared to the wild-type receptor, indicating that the β2AR-T4L construct could still undergo conformational changes on agonist binding. Collectively, these studies show that replacing ICL3 with T4L did not cause significant changes to receptor conformation or flexibility. As was observed in the β2AR/Fab structure, a substantial number of the crystal contacts arise not from the receptor, but from the appended moiety. In the crystal lattice, the T4Ls form symmetry related dimers. On the opposing face of the receptor, a set of lipid-mediated interactions bridge the membrane facing side of TM1 and TM2 to those of a symmetry related partner and includes the palmatoyl groups covalently linked to Cys341. By alternating T4L and lipid mediated interactions, a layer of β2AR/T4L molecules are formed. As these layers are stacked in the crystal, ECL2 and ECL3 form interactions with T4Ls from symmetry related molecules. This appears to have stabilized this region as both the ligand and the extracellular loops are fully resolved (Fig. 2d).

The overall topology of β2AR within the trans-membrane region was nearly identical for both the β2AR/Fab and β2AR/T4L structures, thus confirming that the addition of T4L does not substantially alter the overall receptor conformation. As was observed with the β2AR/Fab structure, the trans-membrane helices pack in a manner similar to rhodopsin, but there are significant deviations on the extracellular side. Relative to rhodopsin, TM1, TM3 and TM4 are pushed away from the center of the receptor bundle while TM5 and TM7 move closer. The net effect of these movements is a more open binding site that can accommodate larger ligands. Within the trans-membrane region, the basic amine from carazolol forms a salt bridge with Asp113$^{3.32}$ from TM3 and hydrogen bonds to Tyr316$^{7.43}$ and Asn312$^{7.39}$. Asn312$^{7.39}$ also hydrogen bonds to the hydroxyl group. The TM5 residue

Ser203$^{5.42}$ forms hydrogen bonds with the NH from the carbazole ring. These interactions are consistent with mutagenesis studies [47, 48], and establish that interactions with Asp113$^{3.32}$, Asn312$^{7.39}$, and Ser203$^{5.42}$ are critical for ligand binding. There are a number of aromatic contacts to the carbazole ring including Phe289$^{6.51}$ and Phe290$^{6.52}$ from TM7 and Tyr199$^{5.38}$ from TM5. Unlike rhodopsin, the β2 ligand does not directly interact with the conserved tryptophan on TM6, Trp286$^{6.48}$, which is the key residue for receptor activation. Instead, it appears that the inverse agonist activity of carazolol arises from packing its carbazole ring against Phe289$^{6.51}$ and Phe290$^{6.52}$ thereby preventing the rotation of Trp286$^{6.48}$ into the activated conformation. The most striking difference between β2AR and rhodopsin is the position and conformation of ECL2. In contrast to the tightly packed β-sheet in rhodopsin, ECL2 in β2 is helical and oriented away from the ligand binding site. Only a phenylalanine side chain, Phe193$^{5.32}$, is projected into the trans-membrane bundle. This results in a solvent exposed binding cavity that allows diffusible ligands to access the site. Within the trans-membrane bundle, the unpaired Asp113 side chain creates a negative electrostatic gradient toward the binding pocket which serves as an attractor for the positively charged ligands.

The β2AR/T4L construct with a stabilizing E122W mutation has given rise to a 2.8Å structure of the partial inverse agonist timolol [49]. The receptor conformation and ligand binding site are largely unchanged from the β2AR/T4L carazolol complex. The hydrogen bonds between Asp113$^{3.32}$, Asn312$^{7.39}$, and Tyr316$^{7.43}$ and the oxypropanolamine are present, but unlike the carazolol structure, there are no direct interactions to TM5. Instead, the morpholino oxygen from timolol forms an indirect interaction with Ser204$^{5.43}$ by first hydrogen bonding with Asn293$^{7.39}$ which then hydrogen bonds to Ser204$^{5.43}$ (Fig. 2e). Also observed in the β2AR/T4L/Timolol complex is a cholesterol binding site located on the membrane side of TM1, TM2, TM3 and TM4 near the intracellular ends of these helices. Although cholesterol was present in the carazolol structure, it was located at the interface between symmetry related monomers which made it impossible to distinguish whether it was a crystallographic artifact or a true binding site. Despite these crystal contacts not being present in the timolol structure due to the anti-parallel orientation of the receptors, two cholesterol molecules were still bound and defined the site. Based on the residues comprising the cholesterol binding site, the authors define a cholesterol consensus motif (CCM) consisting of five residues, [4.39–4.43(R, K]-[4.50(W, Y)]-[4.46(I,V,L)]-[2.41-(F,Y)]. This motif is present in 21% of the Class A family receptors and this subset is predicted to bind cholesterol at the same site as β2AR. A less restrictive definition

that excludes the requirement for an aromatic residue at position 2.41 is found is 44% of the Class A family. Cholesterol has been shown to affect both ligand binding [50] and thermal stability of β2AR, signifying it may be a potential allosteric modulator. The presence of the cholesterol binding site in several other receptors suggests that it may have a similar role in those as well.

One notable difference between the β2AR and rhodopsin structures is the absence of a salt bridge between the conserved D/ER motif (residues 134/135 and 130/131 in rhodopsin and β2, respectively) at the base of TM3 and a glutamic acid (247 and 268) at the base of TM6, known as the "ionic lock" [51] This interaction is present in the inactive rhodopsin structure and thought to be one of the primary interactions that stabilizes its conformation. It is broken in the light activated rhodopsin and β2 structures, suggesting that the β2 conformations more closely resemble an early activated receptor. The higher level of β2 basal activity would also suggest a less stable inactive conformation, consistent with a weaker lock. However, it is difficult to ascertain whether this is a crystallographic artifact caused by the Fab or the insertion of T4L in place of ICL3. Molecular dynamics studies on the β1 and β2 receptors have shown that the ionic lock does form and remains a stable interaction during these simulations, resembling inactive rhodopsin [52, 53]. Additional studies will be required to fully understand the significance of the ionic lock in β2 activation.

The utility of T4L for conformational stabilization has been extended beyond the adrenergic family with the solution of the human A_{2A} adenosine receptor in complex with the antagonist ZM241385 [22]. Antagonist binding affinity was unaffected by the ECL3/T4L swap, but agonist affinity was increased suggesting a bias toward an activated conformation. As would be expected, the helical arrangement is similar to β2AR and rhodopsin, but relative shifts in helical positions result in an rmsd of 2.0–2.5 Å with the largest deviations being at the extracellular face. More pronounced differences between A_{2A} and the other structures are the conformation and position of the extracellular loops, particularly ECL2. Unlike the helical ECL2 in β2AR and the beta sheet in rhodopsin, ECL2 in the A_{2A} receptor lacks a defined secondary structure; instead it adopts a random coil conformation constrained by its connection to TM4 and TM5, and three disulfide linkages to ECL1. One of these ($Cys77^{3.25}$–$Cys166^{5.27}$) is found in most of the Class A family while the other two ($Cys71^{2.69}$–$Cys159^{5.20}$) and ($Cys74^{3.22}$–$Cys146^{4.67}$) are found only in the A_{2A} receptor. These additional disulfide linkages position ECL2 such that it partially occludes the extracellular ends of TM3 and TM5, thereby blocking a portion of the carazolol/retinal ligand binding site and causing a

distinctly different ligand binding mode (Fig. 2a, f). ECL2 forms one wall of the A_{2A} ligand binding site and in doing so disallows a binding mode parallel to the membrane like that of carazolol and retinal. Instead, ZM241385 binds perpendicular to the membrane in a channel parallel to the helical axes formed by TM5, TM6 and TM7 on one side and with contacts to ECL2, TM3 and TM5 on the other. ECL2 contributes one of the primary interactions with the central triazolotriazine core, Phe168$^{5.29}$, anchoring the ligand within binding site (Fig. 1d). The central ring is also in hydrophobic contact with Ile274$^{7.39}$, and forms hydrogen bonds between its exocyclic amine, Glu169$^{5.30}$ and Asn253$^{6.55}$. Asn253$^{6.55}$ also interacts with the oxygen on the furan ring by bridging it to the amine. The furan ring, which is the most deeply buried substituent, projects into a hydrophobic pocket bounded by Leu249$^{6.51}$, His250$^{6.52}$, and Trp246$^{6.48}$. The close proximity to Trp246$^{6.48}$ is particularly significant as this is the highly conserved tryptophan critical to receptor activation. The antagonist activity of ZM241385 can be attributed to its furanyl ring locking the conformation of the tryptophan side chain in the inactive state. Sitting above Trp246$^{6.48}$ are two waters that are hydrogen bonded to each other and the triazolotriazine ring. Although they form no direct interactions with the protein, it is interesting to note that they fill the same cavity between TM2 and TM3 as does the oxypropanolamine moiety of carazolol in β2AR. The phenethylamino group falls onto a hydrophobic patch at the top of TM7 with the phenolic ring itself oriented toward solvent. Given its loose association with the binding site, it would be expected that this position would tolerate broad substitution and this has indeed been observed in the SAR around these compounds [54]. The interactions observed in the structure are also consistent with available mutagenesis data. Mutation of Glu169$^{5.30}$ to alanine results in a loss of both agonist and antagonist affinity [55] and mutation to asparagine rescues antagonist affinity, thus confirming its role as a hydrogen bond acceptor. Mutation of His250$^{6.52}$, Asn253$^{6.55}$, and Ile274$^{7.39}$ to alanine also causes a loss of agonist and antagonist binding affinity [56, 57]. The concord between structural, SAR and mutagenesis data provides a very detailed understanding of antagonist binding to the adenosine A_{2A} receptor.

These structures have been extensively reviewed in the context of receptor function and dynamics. The reader is encouraged to seek out the β2AR reviews from Rosenbaum et al. [58], Shulka et al. [59], Kobilka et al. [60] and Huber et al. [61] for detailed analysis of this system. The reviews from Tiopol et al. [62], Blois et al. [63], and Hanson et al. [64] also provide useful perspectives on this recent work.

4. Utility of the β2AR and A_{2A} Structures in Modeling

4.1. Docking and Scoring to the β2AR Crystal Structure

Almost immediately upon release of the β2AR structure, reports began to appear that validated it as a template for docking and scoring known ligands and also for potentially selecting novel β2 antagonists. The first to appear were companion papers from Topiol et al. [65] and Sabio et al. [66]. In the first manuscript [65], the ability of the docking programs GOLD and Glide-XP to reproduce the carazolol binding mode and to produce credible poses of known β2AR antagonists was assessed. Both were able to generate carazolol poses that overlaid tightly with the crystallographic conformation. Glide-XP was then used to dock a set of known antagonists which shared the carazolol ethanolamine substructure, but replaced the carbazole with different aromatic moieties. The docked poses showed a consistent overlap of the ethanolamine and aromatic portions of these compounds and participated in the same hydrogen bonds as carazolol. The efficiency of this protocol in selecting known beta blockers was also tested by seeding them into a 400,000 compound database and docking the collection. Carazolol, carvedilol, and 11 additional known antagonists were found in the top 100 hits. This validation is extended in the second paper [66], by the inclusion selection and testing of a larger set of compounds from both internal and external databases containing 400,000 and 4,000,000 compounds, respectively. Fifty-six compounds from the internal database were selected and assayed against β2AR with 19 having >35% inhibition at 10 μM and resulted in a hit rate of 36%. Ninety-four were tested from the external database and yielded 17 hits with a 12% hit rate. A diverse set of 320 compounds from the internal database was also assayed with one resulting hit and a hit rate of 0.3%. The authors warned that the very high hit rate for compounds selected from the internal database may have been due to presence of a large number of GPCR ligands, but both screens showed enrichment ratios substantially better than random selection.

Another prospective search reported by Kolb et al. [67], screened the ZINC [68] database of 972,608 commercially available compounds using DOCK. 25 from the top 500 scorers were selected and of these eight measurably displaced 3H-dihydroalprenolol at 20 μM. 6 had >10% displacement and measured K_is for β2 ranging from 0.009 to 3.2 μM. Detailed docking studies suggested that two of the hits mimicked the shape and contacts of carazolol while the remaining four presented unique binding modes which demonstrated that the β2AR crystal structure may be useful in the discovery of novel ligands.

A modified receptor conformation and a set of reference interaction fingerprints [IFP] were utilized by De Graaf et al. [69] to

enhance the ability of GOLD and Surflex to distinguish full and partial agonists from inverse agonists and antagonists. It has been postulated that ligand binding to β2AR is a multi-step process beginning with formation of an ionic bond between the agonist amino group and Asp113$^{3.32}$ followed by packing of the catechol ring against Phe290$^{6.52}$ [70]. A conformational change then allows formation of hydrogen bonds between the catachol hydroxyl groups and serines on TM5; S203$^{5.42}$, S204$^{5.43}$ and S207$^{5.46}$. This hypothesis suggests that the initial receptor conformation for both agonist and antagonist binding should be similar with the receptor then reorganizing to fully engage the agonist De Graaf initially docked the agonist isopreternol into the β2AR/caraozol structure and found that the ethanolamine moiety made identical contacts as carazolol, but the catachol hydroxyls were too far from TM5 to form hydrogen bonds. Changing the rotameric states of S104$^{5.43}$ and S106$^{5.46}$ followed by energy minimization of the complex resulted in a predicted binding mode in which the catachol hydroxyls are in contact with the TM5 serines, as is consistent with experimental data [47, 48]. The new receptor conformation is believed to represent an early-activated form and should be a more relevant template for the selection of agonists. From both structures, binary interaction fingerprints that capture ligand/residue interactions were also derived. A set of 1,016 compounds containing 13 known antagonists/inverse agonists, 13 partial/full agonists, and 980 chemically similar compounds were screened against these receptors using the docking programs Gold and Surflex to assess their performance in selectively identifying ligands from each pharmacological class. The Gold program using the Goldscore scoring function did relatively poorly in retrieving either class of compounds when docked to both receptors, but its performance was significantly enhanced by including IFPs. Surflex enrichment factors for both agonists and antagonists were quite good, but it was unable to discriminate between them. Addition of IFPs to the Surflex scoring function enabled selective scoring. Although the X-ray conformation and the modified early-activated conformation did show some potential to differentiate antagonists and agonists, the authors conclude that the inactive receptor state is an appropriate template for non-specific selection of either class.

The effect of receptor conformation on agonist and antagonist virtual ligand screening (VLS) was also explored Reynolds et al. [71] through construction of a refined β2AR/carazol model and a ligand directed agonist β2AR/isopreterenol model with ICM [72]. Ligand directed β2AR/carazolol crystal structure refinement was carried out in three steps; addition and minimization of hydrogen atoms, biased probability Monte Carlo (BPMC) sidechain and ligand sampling, and minimization of the resulting side chain and ligand conformations. A set of 50 conformations was generated and scored based on predicted binding affinity and

the best scoring conformations were selected. The agonist model was generated by deleting ECL2 and translating TM5 on a vector directed toward TM1 in four increments; 0.5, 1.0, 1.5, and 2.0Å. The translated positions and the original crystallographic coordinates of TM5 were rotated clockwise (as viewed from the extracellular side) 5°, to give a total of 10 TM5 conformations. These models were subjected to BPMC refinement with a starting conformation of isopreteronol generated by flexible alignment with carazolol. The refined models were evaluated in the context of experimental data and ligand binding score. Seven models were selected based on these criteria. A β2AR homology model using the rhodopsin template was also built and included in the VLS comparisons. Two datasets were compiled from the GPCR ligand database (GILDA) [73] as ligand screening test sets: a larger 14,006 compound ligand set containing 347 annotated as β2 binders, and a smaller test set with 954 randomly selected compounds and 15 well known β2AR agonists and antagonists. When screened against the test set, the rhodopsin based homology model, the X-ray structure, and the ligand refined carazolol model provided antagonist enrichment factors of 6.4, 50.9, and 50.9, respectively, within the top scoring 1%. Agonist enrichment factors were substantially lower at 0, 0, and 6.4. The ability of the agonist model to discriminate agonists and antagonists was evident when it was used to screen the larger database as it gave enrichment values of 38.4 and 1.4 for agonists and antagonists, respectively. The X-ray structure and the ligand refined carazolol models showed the opposite behavior with agonist enrichment factors of 1.6 and 13.0 and antagonist enrichment factors of 36.5 and 33.1 for the top scoring 1%. The authors also explored the impact of ECL2 deletion on hit rates and found that it had only a modest effect on enrichment. It was clear that the poor enrichment observed for the rhodopsin model was due primarily to steric issues with the buried conformation of ECL2.

This study has been extended to the construction of more detailed agonist binding models in a subsequent report by Katrich et al. [74]. This work reiterates that point that agonists do fit into the antagonist binding site from the β2AR/carazolol crystal structure and form the same hydrogen bonding network with $Asp113^{3.32}$ and $Asn312^{7.39}$, but cannot form hydrogen bonds with $S203^{5.42}$, $S204^{5.43}$, and $S207^{5.46}$ because of the position of TM5. Displacement of TM5 toward the center of the helical bundle positions the serine side chains such that they can interact with hydrogen bonding substituents on the agonist molecule. Starting from the β2AR/carazolol crystal structure, receptor conformations were generated by Monte Carlo sampling of side chain conformations for residues within 8 Å of carazolol. Additional flexibility was introduced by also sampling backbone torsions for a portion of the ECL2 (191–196) and the residues around the

proline kink in TM5 (197–204), thereby effectively allowing the extracellular side of TM5 to tilt. The best energy conformation of isopreterenol into this family of receptor conformations now exhibited hydrogen bonding to $S203^{5.42}$ and $S207^{5.46}$ which was made possible by a 2Å tilt of TM5 toward the center of the helical bundle. The same strategy was employed in the construction of ten additional agonist and partial agonist models. Within this set, it was found that agonist models with the TM5 tilt were much more predictive of affinity than the corresponding carazolol based models. These results coupled with those from de Graaf et al. [69] suggests that an early activated model, with relatively small conformational changes to the receptor relative to the antagonist bound form, can be used for agonist modeling and virtual screening.

Vilar et al. [75] has also explored the utility of multiple modes in virtual screening. The docked conformations of 94 agonists and antagonists were the basis for constructing multiple structure based, ligand based, and consensus models to predict binding affinities toward β2AR. The ligands were docked into the β2AR/caraozol crystal structure using Glide followed by energy minimization. A top ranking pose for each compound was selected based on energetic and structural considerations. It was assumed that these compounds would all dock in a similar manner, but 27 did not overlap with the consensus binding mode and were eliminated. The remaining 67 were divided into a 53 compound training set and a 14 compound test set for subsequent model validation. Structure based affinity prediction models were built based on these complexes via analysis with Liason (linear interaction energy scoring), MM-GBSA (molecular mechanics with generalized Born salvation energies) and a linear regression model derived from components of the Glide XP scores. The ligand conformations were extracted and the resulting overlay used to build CoMFA/CoMSIA models. A set of 2D descriptor based models were also included. Combinations of the structure based and ligand based models were created using PLS to generate consensus models composed of only structure based terms, of only ligand based terms and of both. Within the structural models, the MM-GBSA model was the most predictive with r^2 values of 0.499 and 0.495 for the test and training sets respectively. The structure based consensus model did not improve performance. The ligand based methods had generally higher r^2 values. For example the CoMFA model had r^2 values of 0.962 and 0.642 for training and test. The ligand based consensus model did show better correlation with r^2 values of 0.942 and 0.757. The models were then subjected to a simulated virtual screen of 344 compounds composed of the 67 β2AR ligands, 52 structurally related dopamine ligands and 225 compounds randomly selected from ChemBank. It was found that both the CoMSIA and

MM-GBSA performed well in retrieving true positives when the pK_i cutoff for activity was set at 7, but that MM-GBSA was more effective at detecting actives when the cutoff was set to 5. The models selected both agonists and antagonists equally well underscoring the utility of the inactive receptor state for virtual screening.

4.2. Comparison of Existing Models to the X-ray Structures

The release of β2AR/carazolol and A_{2A}/ZM241385 has enabled a direct validation of models built by using established GPCR homology modeling protocols with experimental data. Since rhodopsin based modeling is the prevalent methodology, a critical first study was the comparison of a rhodopsin derived β2AR model to the X-ray structure. Costanzi [76] has reported such a study, using two β2AR models built using Modeler [77]. The models differ by their ECL2 conformation, one being derived from the rhodopsin crystallographic coordinates and the second modeled *de novo*. As expected, both models are closely related to the template in the trans-membrane region with backbone RMSDs of 0.32 Å and 0.34 Å, respectively. The ECL2 conformation for the first model also very closely mimics the rhodopsin structure, as it packs on top of the trans-membrane bundle. The de novo ECL2 conformation in the second model is much more solvent exposed, resulting in an open ligand binding cavity as is observed in the X-ray structure. Notably, while the ECL2 position is more closely matches that of β2, the helical motif present in the experimental structure was not predicted. Carazolol was then docked into each of these models using the InducedFit protocol from Schrodinger [78], which allows simultaneous ligand and side chain flexibility during the docking process. For both models, carazolol was oriented properly in the binding site, with the positively charged amine pointing to Asp113$^{3.32}$ and the carbazole moiety oriented toward TM5, but the molecule was positioned much more deeply in the binding site than observed in the crystal structure. In model 1, this can be attributed to steric interactions with ECL2, which closes off the top of the binding site and causes a change in the orientations of several residues lining the binding pocket including Trp109$^{3.28}$, Tyr199$^{5.38}$, Phe290$^{6.52}$, and Tyr316$^{7.43}$. The more open placement of ECL2 in model 2 not only alleviated the steric issues but allowed the binding site residues to assume rotamers equivalent to the β2AR crystal structure with the exception of Phe290$^{6.52}$. The orientation of this residue was apparently the sole reason that carazolol is docked too deeply in model 2 because its conformation places the aromatic side chain away from the binding site, which results in a deeper pocket. Remodeling this residue so that it is oriented into the binding site leads to a docked pose more closely resembling the crystallographic pose with an RMSD of 1.7 Å. To some extent, the success of this study can be

attributed to the similar size of carazolol and retinal and the shape similarity their cognate binding sites. Although the final model reproduced the crystal structure reasonably well, this work does demonstrate the sensitivity of GPCR docking to the conformations of ECL2 and side chains within the binding site. It also provides some insight into the level of accuracy that can be expected from these types of models and the possible challenges that will be encountered when attempting to dock molecules substantially different than the crystallographic ligand into these models.

Ivanov et al. [79] compared a previously published, rhodopsin based adenosine A_{2A} model [80] and a newly constructed β2AR derived model to the A_{2A}/ZM241385 crystal structure. The rhodopsin based model had predicted several residues including, Leu85$^{3.33}$, Phe168$^{5.29}$, Asn181$^{5.42}$, Trp246$^{6.48}$, Leu249$^{6.51}$, Asn253$^{6.55}$, and Ile274$^{7.39}$ to be involved in antagonist binding. The proximity of these residues to ZM241385 was confirmed in the X-ray structure and the model had good overlap of the Cα-atoms in the binding site with an rmsd value of 0.9 Å. The side chain orientations also matched those of the experimental structure within the trans-membrane segments, but there is a much larger deviation for Phe168$^{5.29}$ on ECL2. A binding model of ZM241385 was not reported, but the binding modes of similar compounds were disclosed in the original manuscript [80]. These poses did predict the hydrogen bonding interaction between Asn253$^{6.55}$ and the exocyclic amine on the core heterocycle, but the overall ligand binding mode more closely resembled that of retinal and carazolol rather than the A_{2A} crystallographic orientation. The new work begins by assessing the accuracy of the programs Glide XP and MOE in docking ZM241385 into its crystallographic binding site. Without constraints, MOE oriented the compound properly but failed to accurately place the hydroxyphenethyl moiety and yielded an rmsd of 6.1 Å. The Glide XP pose was inverted with an rmsd of 10.1 Å. In the absence of waters, but with a hydrogen bonding constraint to Asn253, the error in the Glide XP pose was reduced to 2.09 Å. InducedFit docking with the same constraint fared slightly worse with an rmsd of 3.58 Å, but this was due mostly to poor placement of the hydroxyphenyl ring. Inclusion of crystallographic waters produced the most accurate Glide pose, reducing the rmsd to 0.9 Å; however, given the difficulty in placement of water molecules in model structures this is unlikely to be a generally applicable method. The authors also describe agonist and antagonist InducedFit docking into a homology model based on the β2AR/carazolol crystal structure. The model of ZM241385 agreed reasonably well with the crystal although it is placed slightly more deeply into the trans-membrane bundle. A model of the agonist adenosine is also reported, which

predicts the adenine ring to be superimposed with that of ZM241385 and putative hydrogen bonds to Ser277$^{7.42}$ and His728$^{7.43}$. Ser277 has been reported to be critical for agonist activity. Collectively, these results demonstrate the potential of β2AR as a template for agonist and antagonist binding models.

Prior to the release of the A_{2A}/ZM241385 crystal structure, the GPCR modeling community was challenged to predict the structure of this complex in order to provide a true blind assessment of GPCR modeling methodology [81]. Twenty-nine separate groups submitted 206 models (of which only 169 had interpretable ligand structures) that were evaluated based on the accuracy of the predicted protein conformation, ligand binding mode and on the correct number of receptor-ligand contacts. The majority of the models had Cα rmsd values <6 Å with more than half <4 Å, averaging 4.2 ± 0.9 Å. Within the trans-membrane region, Cα rmsd values were even tighter: 2.8 ± 0.5 Å. The accuracy of ligand binding mode prediction varied much more widely with a range of rmsd values from 3 to 39 Å. Interestingly, even among those models with binding site Cα rmsd <4 Å, the range was still 2.8–17.2 Å. Also, few models correctly predicted the 75 possible receptor–ligand contacts. Only one third of the models captured the hydrogen bonding interaction with Asn253$^{6.55}$ and even fewer stack Phe168$^{5.29}$ with the adenine core. The best model, submitted by Costanzi, was a comparative model based on the β2AR/carazolol template with InducedFit ligand docking and incorporated mutagenesis data during interpretation. It showed an extended ligand conformation and the correct orthogonal orientation to the membrane. Although 34 out of 75 contacts are correctly predicted, a few inaccurate side chain placements positioned the ligand too deeply in the binding site resulting in a ligand rmsd of 2.8Å. Comparative modeling based on the β2AR and/or β1AR templates was the most common technique utilized by the participants and was the basis for the top six models. In these models, all had an extended ligand conformation, four had the proper orthogonal ligand orientation, four predicted the hydrogen bonding contact to Asn253$^{6.55}$ and four had an aromatic stack with Phe168$^{5.29}$. Each participant also ranked their models within the set that they submitted, but the correlation of this rank with the accuracy of the model was variable. The best model was ranked first only three times within the top ten submissions.

4.3. Homology Models Using β2AR, β1AR or Adenosine A_{2A} as a Template

Several reports of new GPCR models based on the β2AR, β1AR or Adenosine A_{2A} template structures have appeared in the recent literature. These are summarized in Table 1.

Table 1
Homology models based on the 2AR and Adenosine A_{2A} receptors

Target	Template	References	Result
β3AR	β2AR	[86]	Docking studies used to rationalize potency and selectivity for a set of biphenylacylsulfonamide based β3 agonists.
5-HT_{2A}	β2AR	[87]	Derived a strructural basis for the binding profile of Clozapine and Olanzapine against a panel of 14 receptors. Binding differences ascribed to residue diversity in TM6 and TM7.
5-HT_{2B}			
5-HT_{2C}			
M_1			
M_4			
H_1			
5-HT_{1A}			
5-HT_6			
D_2			
D_3			
D_4			
α_1			
α_2			
Adenosine A_1 Adenosine A_{2A}	β2AR Rhodopsin	[88]	Compared β2AR and rhodopsin based models of antagonist binding to adenosine A_1 and A_{2A} receptors. Interactions in the β2AR based A_{2A}model were deemed to be more stabilizing. Differences were less clear between the A_1 models.
α_{1a} adrenoreceptor	β2AR	[89]	Binding modes of subtype selective α_{1a}-AR antagonists built and refined using simulations in a GBSW implicit membrane model. Models are qualitatively consistent with mutagenesis data.
Chemokine CCR5	β2AR Rhodopsin	[90]	Binding models of CCR5 antagonists anibamine, maraviroc, aplaviroc, TAK779, vicriviroc and SCHC built in both β2AR and rhodopsin based homology models. Both templates seemed reasonable, but the rhodopsin based models are more consistent with mutagenesis data.
Adenosine A_{2B}	β2AR Rhodopsin A_{2A}	[91]	A_{2A} was found to be the best template for A_{2B} modeling. Binding modes for multiple agonists and antagonists are presented and validated with mutagenesis data.
Neurokinin 1	β2AR Rhodopsin	[92]	Assessed the feasibility of a fully automated GPCR docking procedure to generate binding models of CP-96345 in NK-1. Concluded that accurate modeling required multiple templates and refinement with mutagenesis data.

(continued)

Table 1 (continued)

Target	Template	References	Result
cholecystokinin (CCK)	β2AR A_{2A}	[93]	Complexes of two photolabile CCK peptides with the CCK receptor built based on distance constraints derived from photoaffinity labeling and FRET studies. Residues in ECL1 and ECL2 were selectively labeled.
Cannabinoid CB_1	β2AR	[94]	105ns simulation of a CB1 receptor model an explicit POPC membrane. Helical bundle topology remained quite close to the X-ray structure and is stabilized by water-mediated h-bonds, aromatic stacking and receptor-lipid interactions.

5. Conclusions and Future Directions

The next generation of GPCR structure based design has emerged with the release of the β2AR, β1AR and Adenosine A_{2A} crystal structures. The first generation began nearly ten years ago with the rhodopsin structure that gave us the initial high resolution view of GPCR topology and ligand interactions. Because it was the only structure until recently, it became the *de facto* standard template for comparative modeling. These models, while useful for the design of mutagenesis studies and qualitative ligand binding models, lacked true atomic level detail and direct validation. As would be expected within the homologous Class A GPCR family, the subsequent β2AR, β1AR, or Adenosine A_{2A} crystal structures were structurally very similar to rhodopsin, but these new structures showed details not previously predicted. Many had suspected that the rhodopsin β-sheet conformation of ECL2 would not be conserved within the family, but the α-helical structure in β2AR and the intimate contact of Phe168 and ZM241385 in A_{2A} were nonetheless .still surprising revelations to many. These new structures widen our view into the conformational manifold accessible within this protein family and provide a structural context to the body of biochemical and biophysical data that preceded them.

However, these four structures represent only the tip of the iceberg and leave many questions unanswered. For example, modeling has provided some insight into the conformational changes that may occur on agonist binding, but until the first agonist structure is produced these hypotheses will remain unproven. Even then, an agonist structure will only show us a different frozen moment in time; much work remains to fully understand

the conformational changes that occur between the active and inactive states and how this translates into G-protein coupling. It is clear that ECL2 conformation is quite plastic and is dependent on its surrounding receptor-ligand environment. Since it can be a critical determinant for ligand binding, a precise understanding of its potential conformations remains a key challenge. Another question relates to the number of sequences that can be accurately modeled from these templates. Mobarec et al. [82] and Worth et al. [83] have both undertaken sequence based analyses to probe the applicability of these structures for building homology models within the Class A family. They concluded that these structures may not represent suitable templates for a significant fraction of the family and that several additional structures will required for more complete coverage. An understanding of allosteric modulation has not yet been elucidated from the structural work. The cholesterol binding site observed in the β2AR/timolol presents one intriguing possibility for an allosteric binding site, but additional complexes and mutagenesis work will be necessary to define these regions. Finally, since the crystal contacts for each of these structures are derived from an exogenous protein, it is unlikely that the receptor is packing in a physiologically relevant manner. Given the accumulating evidence that many GPCRs function as dimers or higher order aggregates, it will be critical to develop crystallographic methods to fully understand the inter-receptor interactions within these complexes.

Based on the results of GPCR Dock 2008 and the validation studies from Costanzi [76] and Ivanov et al. [79], it is clear that GPCR model building is far from being an automated process. Receptor conformation must be carefully considered, especially with respect to ECL2 and side chain orientations within the transmembrane bundle. Docking studies within the native crystal structures suggest that the current scoring functions are not sufficiently accurate to reliably predict ligand binding modes without additional constraints. These constraints arise from data ancillary to the structural work, usually receptor mutagenesis and ligand SAR, but are critical for the success of any GPCR modeling exercise, particularly when using homology models. Without data of this type, it becomes quite difficult for the modeler to differentiate the correct binding pose from a computationally generated ensemble that may all look reasonable. This is borne out by the results of GPCR Dock 2008 as the most successful entries were those derived from incorporation of mutagenesis data. As can be seen from the references in Table 1, GPCR modeling can lead to a meaningful outcome, but the generation of such models requires sufficient data and the intuition of one practiced in the field.

The current situation is reminiscent of the kinase field in the early 1990s. The first kinase structure, cAMP-dependent protein

kinase [84], revealed its overall fold and the details of ligand interaction within the ATP binding site, but the subsequent structural work has evolved our understanding of kinase activation, loop conformations, ligand binding and allosteric modulation. Ultimately, the large body of kinase crystallographic data and the routine determination of new structure have made structure based design within the kinase family a reality [85]. Some work remains for this to be the case in GPCRs, but it seems likely that, with time, GPCR structure based design will be an integral part of the drug discovery process.

Acknowledgments

The authors would like to thank Dr. Roy Kimura for providing the program to generate the snake plot in Fig. 1 and Dr. Stan Krystek for his input and critical reading of the manuscript.

References

1. Jacoby, E., Bouhelal, R., Gerspacher, M., and Seuwen, K. (2006) The 7 TM G-protein-coupled receptor target family. *Chem. Med. Chem.* **1**, 761–782.
2. Druey, K.M. (2003) Regulators of G protein signalling: potential targets for treatment of allergic inflammatory diseases such as asthma. *Expert Opin. Ther. Targets* **7**, 475–484.
3. Esté, J.A. (2003) Virus entry as a target for anti-HIV intervention. *Curr. Med. Chem.* **10**, 1617–1632.
4. Winzell, M.S. and Ahrén, B. (2007) G-protein-coupled receptors and islet function-implications for treatment of type 2 diabetes. *Pharmacol. Ther.* **116**, 437–448.
5. Thompson, M.D., Cole, D.E., and Jose, P.A. (2008) Pharmacogenomics of G protein-coupled receptor signaling: insights from health and disease. *Methods Mol. Biol.* **448**, 77–107.
6. Hopkins, A.L. and Groom, C.R., (2002) The druggable genome. *Nat. Rev. Drug. Discovery* **1**, 727–730.
7. Horn, F., Bettler, E., Oliveria, L, Campagne, F, Cohen, F.E, and Vriend, G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* **31**, 294–297.
8. Zhang, Y., Ernst, C.A., and Rollins, B.J. (1996) MCP-1: structure/activity analysis. *Methods* **10**, 93–103.
9. Gavrilin, M.A., Gulina, I.V., Kawano, T., Dragan, S., Chakravarti, L., and Kolattukudy, P.E. (2005) Site-directed mutagenesis of CCR2 identified amino acid residues in transmembrane helices 1, 2, and 7 important for MCP-1 binding and biological functions. *Biochem. Biophys. Res. Commun.* **11**, 533–540.
10. Parthier, C., Reedtz-Runge, S., Rudolph, R., Stubbs, M.T. (2009) Passing the baton in class B GPCRs: peptide hormone activation via helix induction? *Trends Biochem. Sci.* **34**, 303–310.
11. Rivier, J., Rivier, C., and Vale, W. (1984) Synthetic competitive antagonists of corticotropin-releasing factor: effect on ACTH secretion in the rat. *Science* **224**, 889–891.
12. Göke, R., Fehmann, H.C., Linn, T., Schmidt, H., Krause, M., Eng, J., and Göke, B. (1993) *J. Biol. Chem.* **268**, 19650–19655.
13. Kunishima, N., Shimada, Y., Tsuji, Y., Sato, T., Yamamoto, M., Kumasaka, T., Nakanishi, S., Jingami, H., and Morikawa, K. (2000) Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor. *Nature* **407**, 971–977.
14. Pin, J.P., Galvez, T., and Prezeau, L. (2003) Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol. Ther.* **98**, 325–354.
15. Schertler, G.F. and Villa, C., Henderson, R. (1993) Projection structure of rhodopsin. *Nature* **362**, 770–772.
16. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le

Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M, and Miyano, M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, **298**, 739–745.
17. Patny, A., Desai, P.V., and Avery, M.A. (2006) Homology modeling of G-protein-coupled receptors and implications in drug design. *Curr. Med. Chem.* **13**, 1667–1691.
18. Fanelli, F. and De Benedetti, P.G. (2005) Computational modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem. Rev.* **105**, 3297–3351.
19. Rasmussen, S.G.F., Choi, H-J, Rosenbaum, D.M., Kobilka, T.S., Thian, F. S., Edwards, P.C., Burghammer, M., Ratnala, V.R.P, Sanishvili, R, Fischetti, R.F., Schertler, G.F.X., Weis, W.I., and Kobilka, B.K. (2007) Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383–387.
20. Cherezov V., Rosenbaum D.M., Hanson M.A., Rasmussen S.G., Thian F.S., Kobilka T.S., Choi H.J., Kuhn P., Weis W.I., Kobilka B.K., and Stevens R.C. (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318**, 1258–1265.
21. Warne, T., Serrano-Vega, M.J., and Baker, J. G., Moukhametzianov, R., Edwards, P.C., Henderson, R., Leslie, A.G., Tate, C.G., Schertler, G.F. (2008) Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **454**, 486–491.
22. Jaakola, V.P., Griffith, M.T., Hanson, M. A., Cherezov, V., Chien, E.Y., Lane, J.R., Ijzerman, A.P., and Stevens, R.C. (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **322**, 1211–1217.
23. Henderson, R. and Schertler, G.F. (1990) The structure of bacteriorhodopsin and its relevance to the visual opsins and other seven-helix G-protein coupled receptors. *Philos.Trans. R. Soc. Lond. B. Biol. Sci.* **326**, 379–389.
24. Underwood, D.J., Strader, C.D., Rivero, R., Patchett, A.A., Greenlee, W., and Prendergast, K. (1994) Structural model of antagonist and agonist binding to the angiotensin II, AT1 subtype, G protein coupled receptor. *Chem. Biol.* **1**, 211–221.
25. Trumpp-Kallmeyer-S, Hoflack, J., Bruinvels, A., and Hibert, M. (1992) Modeling of G-protein-coupled receptors: application to dopamine, adrenaline, serotonin, acetylcholine, and mammalian opsin receptors. *J. Med. Chem.*, **35**, 3448–3462.
26. Pebay-Peyroula, E. Rummel, G, Rosenbusch, J.P., and Landau, E.M. (1997) X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science*, **277**, 1676–1681.
27. Luecke, H, Schobert, B, Richter H-T., Cartailler, J-P., and Lanyi, J.K. (1999) Structure of bacteriorhodopsin at 1.55 a resolution. *J. Mol. Biol.* **291**, 899–911.
28. Ballesteros, J., Palczewski, K. (2001) G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. *Curr. Op. Drug. Discov. Devel.* **4**, 561–574.
29. Baldwin, J.M. (1993) The probable arrangement of the helices in G protein-coupled receptors. *EMBO J.* **12**, 1693–1703.
30. Li, J., Edwards, P., Burghammer, M., Villa, and C., Schertler, G.F.X. (2004) Structure of bovine rhodopsin in a trigonal crystal form. *J. Mol. Biol.* **343**, 1409–1438.
31. Okada, T., Fujiyoshi, Y., Silow, M., Navarro J., Landau, E. M., and Shichida, Y. (2002) Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc. Natl. Acad. Sci.*, **99**, 5982–5987.
32. Okada, T., Sugihara, M., Bondar, A.N., Elstner, M., Entel, P., and Buss,V. (2004) The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure. *J. Mol. Biol.* **342**, 571–583.
33. Murakami, M. and Kouyama, T. (2008) Crystal structure of squid rhodopsin. *Nature*, **453**, 363–367.
34. Nakamichi, H., Buss, V., and Okada, T. (2007) Photoisomerization mechanism of rhodopsin and 9-cis-rhodopsin revealed by X-ray crystallography. *Biophys. J.* **92**, L106-L108.
35. Kimura, S. R., Tebben, A. J., and Langley, D. R. (2008) Expanding GPCR homology model binding sites via a balloon potential: A molecular dynamics refinement approach. *Proteins: Structure, Function, and Bioinformatics* **71**, 1919–1929.
36. Pei, Y., Mercier, R.W., Anday, J.K., Thakur, G. A., Zvonok, A.M., Hurst, D, Reggio, P.H., Janero, D.R., and Makriyannis, A. (2008) Ligand-binding architecture of human CB2 cannabinoid receptor: evidence for receptor subtype-specific binding motif and modeling GPCR activation. *Chem. & Biol.* **11**, 1207–1219.
37. Bizzantz, C., Bernard, P., Hibert, and M., Rognan, D. (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? *Proteins: Structure, Function, and Bioinformatics* **50**, 5–25.

38. Tikhonova, I.G., Sum, C.S., Neumann, S., Engel, S., Raaka, B.M., Costanzi, S., and Gershengorn, M.C. (2008) Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *Biochem.* **51**, 625–633.
39. Carter, P.H. and Tebben A.J. (2009) Chapter 12. The use of receptor homology modeling to facilitate the design of selective chemokine receptor antagonists. *Methods Enzymol.* **461**, 249–279.
40. Ballesteros, J.A. and Weinstein, H. (1995) [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors *Methods Neurosci.* **25**, 366–428.
41. Rosenbaum D.M., Cherezov V., Hanson M.A., Rasmussen S.G., Thian F.S., Kobilka T.S., Choi H.J., Yao X.J., Weis W.I., Stevens R.C., and Kobilka B.K. (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **318**, 1266–1273.
42. Ridge, K.D. and Palczewski, K. (2007) Visual rhodopsin sees the light: structure and mechanism of G protein signaling. *J. Biol. Chem.* **282**, 9297–9301.
43. Gether, U., Ballesteros, J.A., Seifert, R., Sanders-Bush, E., Weinstein, and H., Kobilka, B.K. (1997) Structural instability of a constitutively active G protein-coupled receptor. Agonist-independent activation due to conformational flexibility. *J. Biol. Chem.* **272**, 2587–2590.
44. Serrano-Vega, M.J., Magnani, F., Shibata, Y., and Tate, C.G. (2008) Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. *Proc. Natl. Acad. Sci.*, **105**, 877–882.
45. Kobilka, B.K. (1995) Amino and carboxyl terminal modifications to facilitate the production of a G-protein coupled receptor. *Anal. Biochem.* **231**, 269–271.
46. Granier, S., Kim, S., Shafer, A.M., Ratnala, V.R., Fung, J.J., Zare, R.N., and Kobilka, B. Structure and conformational changes in the C-terminal domain of the beta2-adrenoceptor: insights from fluorescence resonance energy transfer studies. *J. Biol. Chem.* **282**, 13895–13905.
47. Liapakis, G., Ballesteros, J.A., Papachristou, S., Chan, W.C., Chen, X., and Javitch, J.A. (2000) The forgotten serine. A critical role for Ser-2035.42 in ligand binding to and activation of the beta 2-adrenergic receptor. *J. Biol. Chem.* **275** 37779–37788.
48. Strader, C.D., Sigal, I.S., and Dixon, R.A., (1989) Structural basis for beta-adrenergic receptor function. *FASEB J.* **3**, 1825–1832.
49. Hanson M.A., Cherezov V., Griffith M.T., Roth C.B., Jaakola V.P., Chien E.Y., Velasquez J., Kuhn P., and Stevens R.C. (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Structure* **16**, 897–905.
50. Pucadyil, T.J. and Chattopadhyay, A. (2006) Role of chlolesterol in the function and organization of G-protein coupled receptors. *Prog. Lipid Res.* **45**, 295–333.
51. Ballesteros, J.A., Jensen, A.D., Liapakis, G., Rasmussen, S.G., Shi, L., Gether, U., and Javitch, J.A. (2001) Activation of the beta 2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. *J. Biol. Chem.* **276**, 29171–29177.
52. Dror, R.O., Arlow, D.H., Borhani, D.W., Jensen, M.Ø., Piana, S., and Shaw, D.E. (2009) Identification of two distinct inactive conformations of the beta2-adrenergic receptor reconciles structural and biochemical observations. *Proc. Natl. Acad. Sci.* **106**, 4689–4694.
53. Vanni, S., Neri, M., Tavernelli, I., and Rothlisberger, U. (2009) Observation of "ionic lock" formation in molecular dynamics simulations of wild-type β1 and β2 adrenergic receptors. *Biochem.* **48**, 4789–4797.
54. Mantri, M., de Graaf, O., van Veldhoven, J., Göblyös, A., von Frijtag Drabbe Künzel, J.K., Mulder-Krieger, T., Link R., de Vries, H., Beukers, M.W., Brussee, J., and Ijzerman, A.P. (2008) 2-Amino-6-furan-2-yl-4-substituted nicotinonitriles as A2A adenosine receptor antagonists. *J. Med. Chem.* **51**, 4449–4455.
55. Kim, J., Jiang, Q., Glashofer, M., Yehle, S., Wess, J., and Jacobson, K.A. (1996) Glutamate residues in the second extracellular loop of the human A2a adenosine receptor are required for ligand recognition. *Mol. Pharmacol.* **49**, 683–691.
56. Kim, J., Wess, J., van Rhee, A.M., Schöneberg, T., Jacobson, K.A. (1995) Site-directed mutagenesis identifies residues involved in ligand recognition in the human A2a adenosine receptor. *J. Biol. Chem.* **270**, 13987–13997.
57. Jiang, Q., Lee, B.X., Glashofer, M., van Rhee, A.M., and Jacobson, K.A. (1997) Mutagenesis reveals structure-activity parallels between human A2A adenosine receptors and biogenic amine G protein-coupled receptors. *J. Med. Chem.* **40**, 2588–2595.
58. Rosenbaum, D.M., Rasmussen, S.G., and Kobilka, B.K. (2008) The structure and function of G-protein-coupled receptors. *Nature* **459**, 356–363.

59. Shukla, A.K., Sun, J.P., and Lefkowitz, R.J. (2008) Crystallizing thinking about the β2-adrenergic receptor. *Mol. Pharmacol.* **73**, 1333–1338.
60. Kobilka, B. and Schertler, G.F. (2008) New G-protein-coupled receptor crystal structures: insights and limitations. *Trends Pharm. Sci.* **29**, 79–83.
61. Huber, T., Menon, S., and Sakmar, T.P. Structural basis for ligand binding and specificity in adrenergic receptors: implications for GPCR-targeted drug discovery. *Biochem.* **47**, 11013–11023.
62. Topiol, S. and Sabio, M. (2009) X-ray structure breakthroughs in the GPCR transmembrane region. *Biochem. Pharmacol.* **78**, 11–20.
63. Blois, T.M. and Bowie, J.U. (2009) G-protein-coupled receptor structures were not built in a day. *Protein Sci.* **18**, 1335–1342.
64. Hanson, M.A. and Stevens, R.C. (2009) Discovery of new GPCR biology: one receptor structure at a time. *Structure* **17**, 8–14.
65. Topiol, S. and Sabio, M. (2008) Use of the X-ray structure of the Beta2-adrenergic receptor for drug discovery. *Bioorg. Med. Chem. Lett.* **18**, 1598–1602.
66. Sabio, M., Jones, K., and Topiol, S. (2008) *Bioorg. Med. Chem. Lett.* 18, 5391–5395.
67. Kolb, P., Rosenbaum, D.M., Irwin, J.J., Fung, J.J., Kobilka, B.K., and Shoichet, B.K. (2009) *Proc. Natl. Acad. Sci.* **106**, 6843–6848.
68. Irwin, J.J. and Shoichet, B.K., (2005) ZINC – A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* **45**, 177–182.
69. de Graaf, C. and Rognan, D. (2008) Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor. *J. Med. Chem.* **51**, 4978–4985.
70. Kobilka, B.K. and Deupi, X. (2007) Conformational complexity of G-protein-coupled receptors. *Trends Pharmacol. Sci.* **28**, 397–406.
71. Reynolds, K.A., Katritch, V., and Abagyan, R. (2009) Identifying conformational changes of the beta(2) adrenoceptor that enable accurate prediction of ligand/receptor interactions and screening for GPCR modulators. *J. Comput. Aided Mol Des.* **23**, 273–288.
72. ICM-Pro, Molsoft L.L.C., La Jolla, CA. http://www.molsoft.com.
73. Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimoto, R., and Feng, C. (2008) GLIDA: GPCR–ligand database for chemical genomics drug discovery–database and tools update. *Nucleic Acids Res.* **36**, D907–912.
74. Katritch, V., Reynolds, K.A., Cherezov, V., Hanson, M.A., Roth, C.B., Yeager, M., and Abagyan, R. (2009) Analysis of full and partial agonists binding to beta2-adrenergic receptor suggests a role of transmembrane helix V in agonist-specific conformational changes. *J. Mol. Recognit.* **22**, 307–318.
75. Vilar, S., Karpiak, J., and Costanzi, S. (2009) Ligand and structure-based models for the prediction of ligand-receptor affinities and virtual screenings: Development and application to the beta(2)-adrenergic receptor. *J. Comput. Chem.* Epub ahead of print.
76. Costanzi, S. (2008) On the applicability of GPCR homology models to computer-aided drug discovery: a comparison between in silico and crystal structures of the beta2-adrenergic receptor. *J. Med. Chem.* **51**, 2907–2914.
77. Sali, A. and Blundell, T.L. (1993) Comparative protein modeling by satisfaction of spatial constraints. *J. Mol. Biol.* **234**, 779–815.
78. Sherman, W., Day, T., Jacobson, M.P., Friesner, R.A., and Farid R. (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **49**, 534–553.
79. Ivanov, A.A., Barak, D., and Jacobson, K. A. (2009) Evaluation of homology modeling of G-protein-coupled receptors in light of the A_{2A} adenosine receptor crystallographic structure. *J. Med. Chem.* **52**, 3284–3292.
80. Kim, S.K., Gao, Z.G., Van Rompaey, P., Gross, A.S., Chen, A., Van Calenbergh, S., and Jacobson, K.A. (2003) Modeling the adenosine receptors: comparison of the binding domains of A2A agonists and antagonists. *J. Med. Chem.* **46**, 4847–4859.
81. Michino, M. and Abola, E., GPCR Dock 2008 participants, Brooks, C.L., 3rd, Dixon, J.S., Moult, J., Stevens, R.C. (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nat. Rev. Drug. Disc.* **8**, 455–463.
82. Mobarec, J.C., Sanchez, R., and Filizola, M. (2009) Modern homology modeling of g-protein coupled receptors: which structural template to use? *J. Med. Chem.* **52**, 5207–5216.
83. Worth, C.L., Kleinau, G., and Krause, G. (2009) Comparative sequence and structural analyses of G-protein-coupled receptor crystal structures and implications for molecular models. *PLoS One* **4**, e7011.
84. Knighton, D.R., Zheng, J.H., Ten Eyck, L.F., Xuong, N.H., Taylor, S.S., and Sowadski, J.M. (1991) Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* **253**, 414–420.
85. Johnson, L.N. (2009) Protein kinase inhibitors: contributions from structure to clinical compounds. *Q. Rev. Biophys.* **42**, 1–40.

86. Hattori K, Orita M, Toda S, Imanishi M, Itou S, Nakajima Y, Tanabe D, Washizuka, K., Araki, T., Sakurai, M., Matsui, S., Imamura, E., Ueshima, K., Yamamoto, T., Yamamoto, N., Ishikawa, H., Nakano, K., Unami, N., Hamada, K., Matsumura, Y., and Takamura, F. (2009) Discovery of highly potent and selective biphenylacylsulfonamide-based beta3-adrenergic receptor agonists and molecular modeling based on the solved X-ray structure of the beta2-adrenergic receptor: part 6. *Bioorg. Med. Chem. Lett.* **19**, 4679–4683.
87. Selent, J., López, L., Sanz, F., and Pastor, M. (2008) Multi-receptor binding profile of clozapine and olanzapine: a structural study based on the new beta2 adrenergic receptor template. *Chem. Med. Chem.* **3**, 1194–1198.
88. Yuzlenko, O. and Kieć-Kononowicz, K. (2009) Molecular modeling of A1 and A2A adenosine receptors: comparison of rhodopsin- and beta2-adrenergic-based homology models through the docking studies. *J. Comput. Chem.* **30**, 14–32.
89. Li, M., Fang, H., Du, L., Xia, L., and Wang, B. (2008) Computational studies of the binding site of alpha1A-adrenoceptor antagonists. *J. Mol. Model.*, **14**, 957–966.
90. Li, G., Haney, K.M., Kellogg, G.E., and Zhang, Y. (2009) Comparative docking study of anibamine as the first natural product CCR5 antagonist in CCR5 homology models. *J. Chem. Inf. Model.* **49**, 120–132.
91. Sherbiny, F.F., Schiedel, A.C., Maaß, A., and Müller, C.E. (2009) Homology modelling of the human adenosine A_{2B} receptor based on X-ray structures of bovine rhodopsin, the β_2-adrenergic receptor and the human adenosine A_{2A} receptor. *J. Comput. Aided. Mol. Des.* Epub ahead of print.
92. Kneissl, B., Leonhardt, B., Hildebrandt, A., and Tautermann, C.S. (2009) Revisiting automated G-protein coupled receptor modeling: the benefit of additional template structures for a neurokinin-1 receptor model. *J. Med. Chem.* **52**, 3166–3173.
93. Dong, M., Lam, P.C., Pinon, D.I., Abagyan, R., and Miller, L.J. (2009) Elucidation of the molecular basis of cholecystokinin Peptide docking to its receptor using site-specific intrinsic photoaffinity labeling and molecular modeling. *Biochem.* **48**, 5303–5312.
94. Shim J.Y. (2009) Transmembrane helical domain of the cannabinoid CB1 receptor. *Biophys. J.* **96**, 3251–3262.

Chapter 16

Methods for Combinatorial and Parallel Library Design

Dora M. Schnur, Brett R. Beno, Andrew J. Tebben, and Cullen Cavallaro

Abstract

Diversity has historically played a critical role in design of combinatorial libraries, screening sets and corporate collections for lead discovery. Large library design dominated the field in the 1990s with methods ranging anywhere from purely arbitrary through property based reagent selection to product based approaches. In recent years, however, there has been a downward trend in library size. This was due to increased information about the desirable targets gleaned from the genomics revolution and to the ever growing availability of target protein structures from crystallography and homology modeling. Creation of libraries directed toward families of receptors such as GPCRs, kinases, nuclear hormone receptors, proteases, etc., replaced the generation of libraries based primarily on diversity while single target focused library design has remained an important objective. Concurrently, computing grids and cpu clusters have facilitated the development of structure based tools that screen hundreds of thousands of molecules. Smaller "smarter" combinatorial and focused parallel libraries replaced those early un-focused large libraries in the twenty-first century drug design paradigm. While diversity still plays a role in lead discovery, the focus of current library design methods has shifted to receptor based methods, scaffold hopping/bio-isostere searching, and a much needed emphasis on synthetic feasibility. Methods such as "privileged substructures based design" and pharmacophore based design still are important methods for parallel and small combinatorial library design. This chapter discusses some of the possible design methods and presents examples where they are available.

Key words: 3D pharmacophores, Library design, Combinatorial library, Structure-based design, Target family library, Gene family, Target family knowledge database, Cell-based library design methods, BCUTs, Privileged substructure, DiverseSolutions, ClassPharmer™, Parallel library design, Scaffold hopping, Diversity, Bio-isosteres, Fragment based design, CombiGLIDE, DOCK, FLAP, SHOP, ROCS, EON, CombiDOCK, OptiDOCK

1. Introduction

In the early 1990s, the pharmaceutical industry was utterly bedazzled by the combinatorial chemistry revolution. The possibility of combining automation, high throughput screening, solid phase synthesis, and sophisticated methods for identifying

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_16, © Springer Science+Business Media, LLC 2011

compounds in mixtures [1] resulted in an era of huge combinatorial endeavors that ranged from tens to hundreds of thousands of compounds per library [2–4]. In the case of peptide libraries, their size even swelled into the millions [4–6]. Initially, library design efforts focused on the production of these large numbers of products to augment high-throughput screening (HTS) decks [7–14]. These libraries, whose sizes dwarfed the typical corporate compound collections of the time, were expected to yield unprecedented numbers of exciting, new leads. The optimistic belief that high-throughput screening of these huge libraries might yield new drugs as readily as Athena sprang from forth full grown from the head of Zeus proved unrealistic [15]. Early design efforts for some of these mammoth libraries were limited at best. By mid-decade, serious design efforts began through application of various types of statistical experimental design [Design of Experiments (DoE)] [16–19] or cell-based approaches [20, 21]. A new criterion that reached beyond mere numbers, diversity, became the design paradigm of the decade. By its very nature, diversity is a qualitative measure and can only be determined in a relative sense. It is necessary to define a specific property or property set for even relative quantification. As has often been stated in analogy to beauty, "diversity is in the eye of the beholder" and although many methods were developed for diverse library design [12], the varying definitions of diversity still remained as diverse as the concept is relative [22].

In order to create a frame of reference for diversity, the concept of "chemistry space" [23] emerged. In essence, a chemistry space is a multi-dimensional matrix defined by the properties of interest. These properties could be based upon various 2D or 3D structural descriptors of the reagents or of the products, upon calculated properties or upon quantitative structure activity relationships (QSAR) models.

Even if one imposes a specific definition of diversity on an ultra large library, it is exceeding difficult to design such a library within the constraints of a full combinatorial matrix. Large numbers of compounds with similar properties tend to result because of the nature of chemical reactivity. Synthetic reality imposes limitations on the diversity of the reagents that can be used if one expects to actually get products from a given reaction. Typically, a large combinatorial library has a very densely populated core region with significant population at the edges of design space and large regions that are either very sparsely populated or are devoid of compounds. The net result is a library that is much less diverse than expected. It might be useful for extracting limited structure activity relationship (SAR) – if the compounds are actually isolated and their HTS results confirmed – but the overall information content of such a library is low relative to the cost of reagents for synthesis and screening, even in HTS formats.

The compounds tend to be of the "methyl, ethyl, butyl, futile" variety – a non-optimal scenario for general lead finding. In addition, a significant percentage of the compounds lie outside of drug space because they derived from undesirable combinations of reagents that arose in the full combinatorial synthesis matrix. Making such compounds is a waste of resources, but to design them out of the library further limits its diversity in the property space of choice.

Several solutions to the ultra-large library design issues were found. One was the design of sparse matrices or incomplete combinatorials [24, 25]. This required division of the combinatorial matrix into sub-matrices to allow greater control in the selection of desirable reagents and for the omission of undesirable ones. While this reduced overall synthetic efficiency, it allowed both increased diversity and the elimination of at least some of the extremely non-drug-like products. In actual practice, the sub-matrix solution could be translated into a series of smaller libraries that were not necessarily constrained to have identical reaction conditions. Another possible solution was simply to make smaller libraries and put more effort into property based design.

The advent of these smaller more "flexibly designed" libraries that consisted of hundreds or a few thousand compounds rather than the tens or hundreds of thousands of compounds, allowed the incorporation of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties [26, 27], or for focus on a receptor target family or even for the design of a library for a specific target. These single target-focused libraries remain a key component of many drug discovery programs, and are useful in both "hit to lead" and "lead optimization" contexts, while creation of "drug-like" and target family libraries largely replaced generation of purely diversity based libraries.

From the late 1990s and into the early twenty-first century, combinatorial chemistry not only trended away from ultra-large diverse libraries to medium sized (1–10 K) libraries, but also to ones that employed the Lipinski's "Rule of Five" [28], and subsequent analyses of "drug-likeness" [29–31]. Screening of large libraries of high molecular weight, greasy (high logP) compounds had yielded non-progressible hits and led to the realization that a definition of "drug-likeness" was essential for the design process. As corporate collections grew in size, the expense of HTS increased and the need to avoid the waste of screening reagents on non-progressible leads arose.

Meanwhile, using variations upon Evan's original definition of privileged substructure [32, 33] in addition to other methods, designs of target family libraries, particularly for G-protein coupled receptors (GPCRs) [34–36] but also kinases [37, 38], nuclear hormone receptors (NHRs) [39], and enzymes such as serine

proteases [40] became increasingly popular. This trend was greatly aided by the classification of known drugs according to their gene ontology by Schuffenhauer [41] and prompted the rise of commercial knowledge databases [42–44] of target family ligand information.

These target family libraries fell between the two extremes of diverse and target focused libraries [45]. These were combinatorial libraries in which the products were biased towards multiple members of families of related receptors or enzymes, rather than individual targets [41]. The basis for the design and synthesis of target family combinatorial libraries was the observation that receptors and enzymes that belonged to the same functional family (e.g. kinases, Class I GPCRs) often shared similar steric and electronic features in their binding/active sites. Kinases, which have been shown to have highly similar active sites, provide one example of this [46]. Another example is provided by the Family A GPCR receptors [47]. Several of these are presumed to require aromatic rings and basic moieties in their ligands [48]. Identification and exploitation of these intra-family similarities afforded the opportunity to design and synthesize target-class combinatorial libraries in which the products contained features that were complementary to the common motifs found in the binding/active sites of the receptors or enzymes comprising the particular target family.

Target family libraries appeared well suited for augmenting HTS decks with "drug-like" compounds designed to include key receptor/enzyme binding features. In addition, they have proven valuable in focused screening campaigns where compounds are assayed against multiple targets from the same family (presumably, but not necessarily, the family for which the library was designed). Another intended potential application of target family libraries was the de-orphaning of biological targets of unknown function.

While "drug like" and target family libraries are suitable for lead-finding and perhaps for initial chemistry space mapping in lead optimization for a particular target or group of targets, smaller combinatorial libraries of a few hundred compounds and parallel libraries (ranging from dozens to at most a few hundred compounds) have come to play a major role in lead optimization. Because these libraries contain limited numbers of compounds, it is essential to maximize information about structure activity relationships in the design. Experimental design methods that received limited acceptance for large library design are being re-examined [49]. These design methods may also incorporate QSAR [50] and/or ADMET [30] property models.

Another trend which has significantly impacted library design is that of "Small is Beautiful" or "lead-like" libraries [51, 52]. Lipinski [28] has often pointed out that combinatorial chemistry

produced large numbers of increasingly large, lipophilic compounds. Molecular weight tended to be greater than 500 and often greater than 700 with ClogP [53, 54] values greater than 5 or even higher. Such compounds are poor starting points for medicinal chemistry because these properties generally continue to increase as potency and selectivity are optimized. As a result, these compounds leave little room for lead optimization before reaching prohibitively high weights or logP values. Oprea and others [52, 55–57] have advocated a "lead-like" or fragment-based approach that starts with design of libraries of smaller compounds which are screened for hits that can be subsequently optimized for potency through parallel library design and traditional medicinal chemistry methods.

As collections of combinatorially derived molecules have grown, intellectual property and novelty have become major design considerations. As a result, bio-isosteric replacements and scaffold-hopping tools are now major elements of current library design methodology.

1.1. Computational Requirements for Target Family and Focused Library Design

Regardless of which library design paradigm is employed, computational methods which can identify the interactions responsible for ligand–receptor binding and/or the molecular features needed to form these interactions are required for the library design process. The preferred method is direct examination of high-resolution crystal structures of the enzymes or receptors co-crystallized with ligands. This provides information regarding the explicit interactions relevant to ligand binding, and also allows the shapes of the binding/active sites to be compared. Given this type of data, docking is an extremely powerful computational tool that can be utilized for library design. Structure based methods will be considered in detail.

Unfortunately, it is often the case that crystal structures are not available for targets of interest. This limitation is especially acute for GPCRs, which are the biological targets of as many as 50% of recently launched drugs [48]. The paucity of structural data for GPCR targets is offset, at least partially, by the large amount of data available for classes of ligands, which bind to these targets. Recently, several GPCR X-ray crystal structures have become available [58–63] and the use of homology modeling has become commonplace for GPCRs [64, 65] and for other targets [66–70] as well.

Computational models relating molecular structure and/or properties to biological activity are required for the design of both target-focused and target family combinatorial libraries based on known active ligands. These models are developed from descriptors, which encode information about molecular properties and structure. Many different descriptor types ranging from simple physicochemical properties (e.g. molecular weight, cLogP [71],

rotatable bond count) to 2D descriptors based on molecular connection tables (e.g. atom pairs [72], Daylight Fingerprints [73] to 3D pharmacophores [45, 74, 75]) and 3D property derived BCUTS [76, 77] have been utilized for library design purposes. The use of pharmacophores and of BCUTs will be discussed in detail.

1.2. Design Considerations for Libraries of All Types and Sizes

There are two basic paradigms for the design of target family combinatorial libraries. These are applicable for target family libraries composed of compounds "cherry-picked" from a larger screening set, as well as combinatorially synthesized libraries. The following descriptions assume an ultimate library design goal of a combinatorially synthesized, target family library of 5,000 compounds for a target family composed of five members.

In the first design approach, the common features required for ligand binding to all (or many) of the target family members are identified, then used to derive a model for selecting library products for a 5,000-compound library based on one or more combinatorial templates. In the second approach, five 1,000-compound combinatorial libraries are designed where each library is directed at a single member of the target family using target-specific computational models. These five focused libraries are then combined to form the large 5,000-compound target family library. The amount of effort required (both computational and synthetic) for the first method is much less than that required for the second approach. However, intuitively, a library of this type would be expected to provide weak, non-selective hits against members of the target family since the model used to design the library emphasizes their similarities, rather than the differences which are responsible for ligand specificity.

The second approach, although requiring more effort, has the potential to provide some combinatorial products which are potent against the individual targets, and other products which bind to members of the target family which were not explicitly considered in the design effort. Potency against the individual targets depends upon the quality of the individual focused designs that make up the target family library. However, even the most successful focused library designs provide product sets in which only a fraction of the products bind to their intended targets. Combinatorial products which as a result of some unfavorable interaction, do not bind to the target for which they were designed, may bind to a related target where the unfavorable interaction is absent. This latter approach for target family library design has the added benefit of providing "more shots on goal."

There are no published examples directly comparing these two paradigms for target family combinatorial library design. However, in practical experience, it is often the case that the

compounds which were designed during lead optimization phases of projects focused on particular biological targets are identified as hits in HTS assays run for other targets within the same family. By extension, it is reasonable to expect that compounds from combinatorial libraries designed against one member of a target family, may also bind to other enzymes/receptors within that same family.

When designing a combinatorial or parallel library, it is essential to consider the intended use of the library. The library size and makeup of a general screening library are generally different from that of a target family library, which is, in turn, different from a library intended for a single target or for optimization of a lead. Clearly, the largest library type is the general screening or corporate deck enhancement library. These are most likely to be based on some definition of molecular diversity, but the nature of that diversity will be dependent on whether the library centers on a single scaffold or around a diverse set of scaffolds. The former is mostly likely to be a medium size library on the order of 1,000 compounds or less whereas the latter may range from one to tens of thousands. Similarly, a target family library is likely to be diverse but focused around a set of so-called privileged substructures or around a set of pharmacophores. Typically, these libraries are designed to be in the range of 5–30,000 compounds. Again, if they center on a single scaffold or privileged substructure they will be an order of magnitude smaller. This will depend on whether the library is intended for general target family screening, including de-orphaning of receptors, or if the intent is to enhance a corporate collection. If such libraries are focused on a few receptors the size will obviously be reduced. Libraries for lead optimization generally tend to be on the order of one thousand compounds or less if they are combinatorial. While these libraries may have elements of diversity, they should be designed to yield SAR information if at all possible. More commonly for lead optimization, the library is synthesized in parallel format and size ranges from tens to at most a few hundred compounds. These are extremely SAR driven libraries and often employ classical SAR approaches such as that of Hansch and Wilson [78] or Topliss [79]. Alternately, they may be designed using statistical experimental design approaches [20, 49, 80–82] as well.

Whether a library is based on diversity or focused on one or more targets or scaffolds, it is also necessary to consider what constraints will be imposed by ADMET properties. Commonly, cutoffs for molecular weight and logP that are based on application of Lipinski's rules are employed for all library types. Frequently, the cutoff of MW < 500 and logP < 5 are relaxed to MW < 600 or sometimes 700 and logP < 6 or 8 for full combinatorial libraries. There are, of course, two ways of applying these cutoffs; either the cutoffs are applied to the products in a virtual library or the initial reagent lists are culled. The former is

preferred since culling reagents results by molecular weight may result in the loss of interesting products that arise from combinations of very small and large reagents [25]. Other ADMET related properties may also be applied to the library design, particularly if solubility, adsorption, blood brain barrier (BBB) or other ADMET based QSAR models [50] are available. In general, such models are not applied to general screening libraries, but their use is becoming increasingly important for lead optimization libraries [28, 83, 84].

Library design methodologies, whether diverse or focused, fall into two basic categories: reagent based selection methods and product based selection methods. While product based approaches are clearly the optimal choices [85], practical considerations dictate that synthetic chemists work in reaction space because they are concerned with selection of reagents from commercial or in-house custom sources. Additionally, virtual libraries of combinatorial products even for single scaffolds may number in the millions or billions of compounds. Methods such as those employed in Pearlman's DiverseSolutions [24, 86] and Schrodinger's CombiGLIDE [87] provide a compromise of "reagent biased" product selection, thereby gaining the advantages of both methods.

For small or parallel libraries intended for a single target, combinatorial docking may be a viable focused design option – provided a suitable receptor structure and scoring function exists. The size of the virtual library will also be a determining factor unless the reagent biased approach (such as that underlying CombiGLIDE) is employed. When suitable receptors structure are unavailable, ligand-based methods such as pharmacophores or BCUT chemistry spaces with defined active cells typically used for larger libraries may be employed, in addition to QSAR, Hansch or Topliss based methods.

All libraries are constrained by available reagents and the synthetic methods employed. An essential part of virtual library generation and, if required, full library enumeration, is preparation of the reagent lists. Known non-reactive and multi-reactive reagents have to be removed, in addition to any compounds in the list that inadvertently do not contain the appropriate reactive moiety. Many enumeration tools for virtual libraries have been developed. A representative example and one of the most thorough was Tripos Benchware which was developed in the laboratory of Robert Pearlman and formerly known as Optive Benchware or Librarymaker/ LibraryDesigner [88]. Of special utility in this program was an extensive list of undesirable/ desirable fragments that were used for filtering synthetically feasible reagents. The program also allowed user control of multiple occurrences of the reactive center.

2. Methods

Most of the available computational methods for library design can be applied to target family and focused or parallel libraries. In this chapter, rather than striving for a complete review of all possible methods, we will focus on selected significant methodologies, both historical and current. Among them are 3D pharmacophore descriptor based design applications, privileged substructure methods, cell-based design methods, and structure-based methods. Where design examples exist, they will be discussed. Newer tools such as those for scaffold-hopping and bio-isostere searching will also be addressed in the context of library design.

2.1. Descriptor and Property Based Diversity Methods

2.1.1. Properties, Principle Component Analysis, Clustering, and Design of Experiments

The use of property descriptors was among the earliest methods applied to the creation of diverse compound libraries. Since many ligand derived properties such as molecular weight, logP, fingerprints, atom pairs [89], pharmacophores [90], MolconnZ [91] topological, and E-state descriptors [92] are inter-correlated, it is necessary to use methods such as principal component analysis (PCA), to create orthogonal axes. Once these axes have been defined it is possible to use partial least squares (PLS) regression to define the property space, then distance based selection of maximally diverse compounds is performed. This type of methodology is exemplified by the tools developed by Waldman et al. [93] for Cerius2 [94]. Later modifications of the method imposed ADMET constraints such as Lipinski's "Rule of Five" to constrain compound selection to "drug space" [95, 96].

Another early property-based approach drew upon classical experimental designs, such as full-factorial [20, 21, 82] and D-optimal [16–18, 80] designs, to sample diversity space of either the ligands and/or the products. Statistical experimental design is commonly used for process chemistry and other applications where multiple variables have to be optimized simultaneously. The application to analog synthesis was proposed by both Austel [81] and by Brannigan et al. [19–21, 82]. Using a two level, three parameter full factorial as an example, one chooses a set of properties and assigns a threshold cutoff to each. One can now assign each molecule to a "bin" according to its properties thusly:

	Prop1	Prop2	Prop3
Moleculeset1	+	+	+
Moleculeset2	+	+	−
Moleculeset3	+	−	+

(continued)

(continued)

Moleculeset4	–	+	+
Moleculeset5	–	+	–
Moleculeset6	–	–	+
Moleculeset7	–	–	–

Selection from the bins may be performed manually by a chemist if a relatively small set of reagents is being sampled (R-group selection) or in an automated fashion by any of a variety of possible methods for products or large reagent sets. In principle, this method allows the user to constrain diversity space by omitting bins that represent non-drug-like regions of the property space. It also allows the user to select additional similar compounds or reagents based upon the biological response found for the library or array. In practice, it is necessary to expand the number of "threshold cutoffs" to allow finer sampling of the properties of interest rather than perform repeated iterative selections for combinatorial libraries. This leads logically to cell based diversity analysis, particularly for dealing with large virtual libraries, and which will be discussed below.

The statistical experimental design (DoE) methods, however, are well suited to the design of array libraries for SAR exploration. In this scenario, the chemist designs a mono-variate or "one-by" library. Since only one position is being optimized, the compound selection is easily done in "reagent space" rather than "product space". While the array is mono-variate in the sense that only one R-position is being evaluated, it is in fact multivariate in properties of interest to the chemist: steric bulk or hydrophobicity, hydrogen bonding, polar surface area, etc. Thus, DoE-based approaches provide a reasonable alternative to or complement the traditional medicinal chemistry approaches of Hansch [78] and Topliss [79]. It must be pointed out, however, that the results of such arrays should be used with caution if the chemist plans on a "best-of-best" approach to optimizing R-groups. It does not necessarily follow that the results of varying different positions should be additive. As shown in Fig. 1, optimal combination of R-groups could be easily missed if a combinatorial approach is not applied, either via a combinatorial library or through synthesis of multiple arrays. Wold [16–18, 97] has frequently pointed out that multivariate designs find optimal experimental conditions and compounds missed by traditional one variable (or R-group) at a time (OVAT) approaches historically used for medicinal chemistry optimization. Wold developed tools, including MODDE [98], for compound design in multivariate scenarios. It has been demonstrated that this approach can uncover non-additive SAR

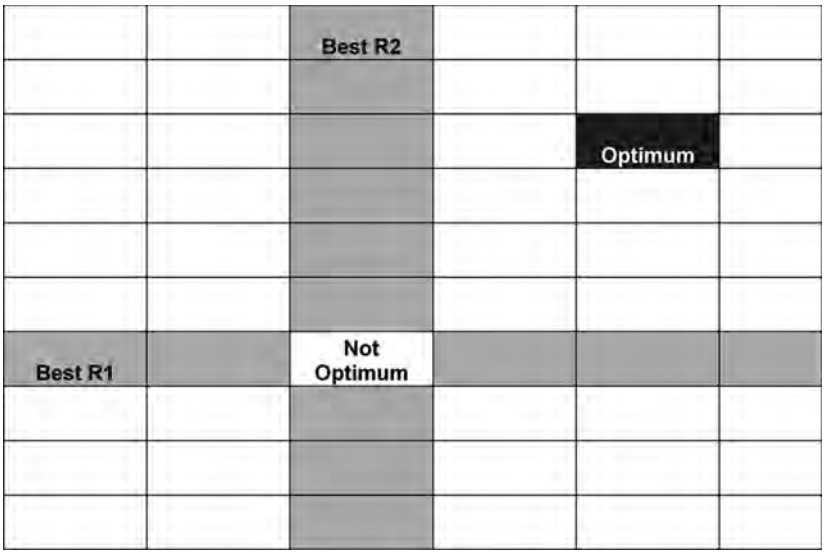


Fig. 1. The dangers of mono-variation when using parallel libraries to find the optimum compound.

effects that would normally be missed by the traditional OVAT [97, 99] design. Clearly, these methods provide a huge advantage over OVAT library approaches for small, diverse libraries. Their use in the design of parallel and combinatorial libraries has been on the increase [49]. Naturally, scenarios where variation of only a single substituent position may appear additive sometimes arise. For serine proteases, for example, moieties such as hydroxamates in matrix metalloproteinases (MMPs) [100] and the benzamidine for serine proteases [101] dominate ligand binding interactions in the receptor. In these cases, OVAT optimization of groups extending into the other pockets can be effective. Such an approach is much more risky in scenarios where the binding mode is unknown and no "warhead" moiety is known to control ligand binding to the receptor.

Other property based methods for diverse selection such as fingerprints, distance-based clustering, Kohenen Maps and spanning trees have been reviewed at length [12]. These methods will not be described here.

2.1.2. Cell-Based Methods for Diverse and Focused Libraries

While the experimental design strategies described above do, strictly speaking, fall into the category of cell-based methods, the most commonly applied method is exemplified by DiverseSolutions [86] from the Pearlman group at the University of Texas at Austin. Much of the library design functionality, particularly for focused libraries, was subsequently implemented as LibraryDesigner in Tripos Benchware [88].

The cell-based method divides each axis of a multidimensional property space into bins and thereby divides the

space into hyper-cubes or cells. Molecules of interest such as known ligands of individual targets or target families can thus be associated with the cells they occupy. One of the difficulties with distance-based methods such as the clustering of Daylight [73] fingerprints is that the result is dependent on pair-wise comparisons of all the structures in the virtual library. Addition or deletion of compounds to the original set may change the number, size and/or membership of the clusters. By contrast, cell based property spaces allow compound addition and deletion without alteration of the chemistry space. This allows not only comparison of libraries in a chemistry space (Fig. 2) but also the definition of hot spot regions in the chemistry space if in fact actives do cluster.

One of the basic assumptions regarding cell based methods is that actives possess common properties which cause them to cluster in a "chemistry space". Minimally, unknown actives should be found in the same region of chemistry space as the known ligands. If the descriptors are truly meaningful, the most active compounds should be clustered more tightly than less active compounds. Clearly, the validity of this assumption about "nearest neighbors" is dependent on the relevance of the chemistry space descriptors to ligand–receptor binding. Given that assumption, this method easily lends itself to ligand based target family or to focused design if a knowledge database of the target ligands is available to define a target family space that contains regions where ligands for specific target/s cluster.

DiverseSolutions [86], employs a unique set of descriptors, BCUTS [76, 102] that are based upon both connectivity and atomic properties such as charge, polarizability and hydrogen bonding properties that appear to correlate with ligand receptor binding activity [20, 21]. Once BCUT descriptors of various

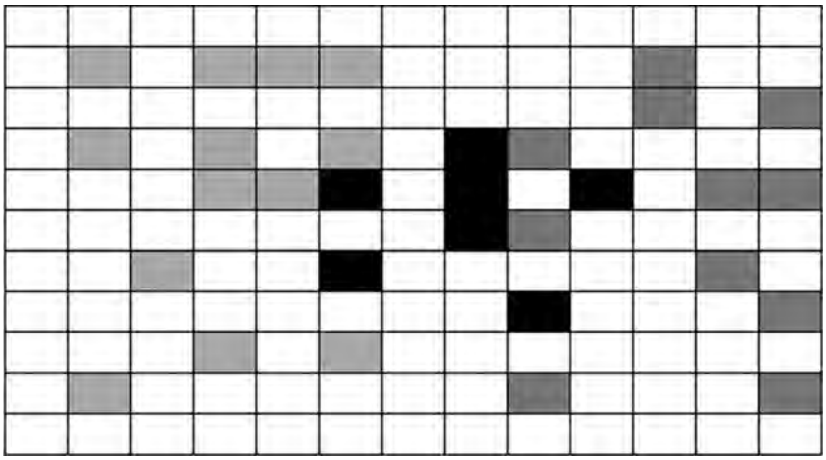


Fig. 2. 2D representation of two compound sets in a cell based chemistry space. Cells occupied by set one in *light gray*, by set two in *dark grey*, and by both sets in *black*. Empty cells are *white*.

"flavors" are calculated for a fully enumerated virtual library, the optimal chemistry space axes that describe the library are determined by finding an orthogonal set of four to eight descriptors that distribute the compounds evenly across the space according to a CHI-squared algorithm [76, 103]. The space is then partitioned into cells or hypercubes by subdividing these axes.

An example is shown (Fig. 3) for an ion channel target. Channel openers and blockers are differentiated in the cell-based space derived from an optimally derived diversity space for the entire combinatorial library that contained them. Since the diversity space in this case has four dimensions and the plots are 3D, it is possible to observe that some subsets of the descriptors seem to cluster and separate the blockers and openers better than others. Qualitative observations that BCUTs do appear to correlate with ligand receptor binding and activity [20, 21, 77] contributed to the development and implementation of the concept of receptor relevant subspaces [77] in the DiverseSolutions software package. This concept is so integral to Pearlman's approaches that it needs to be discussed in some detail. Observations of the type of

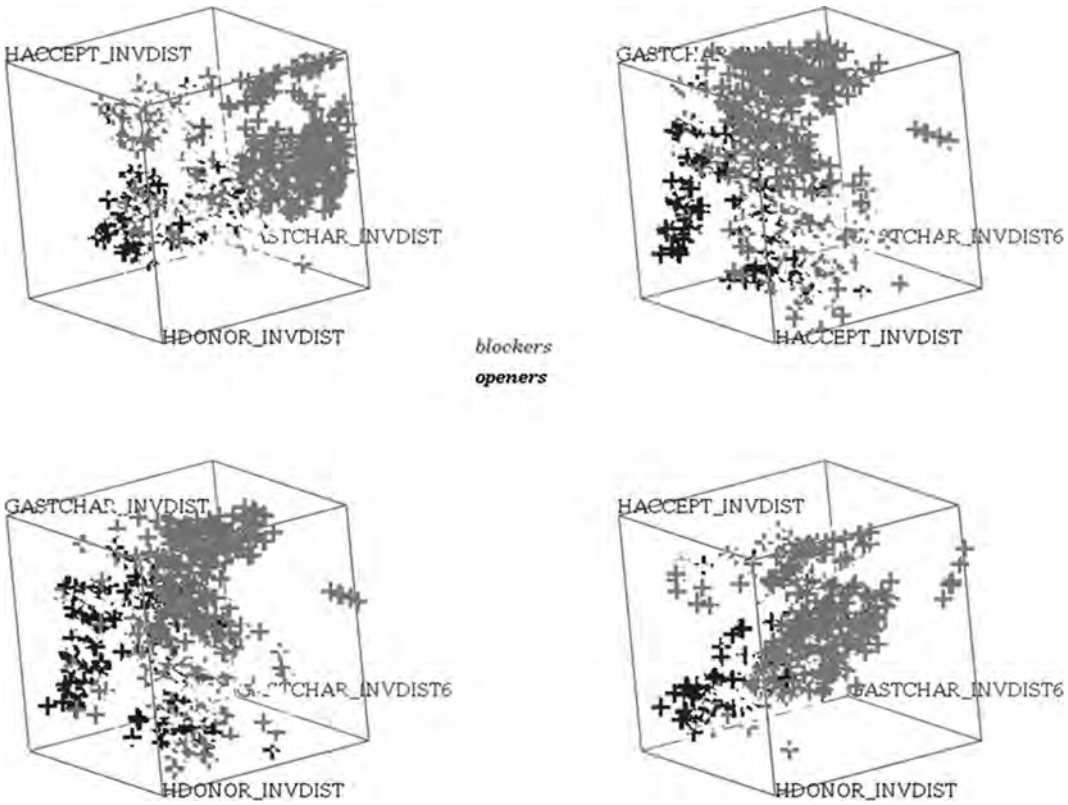


Fig. 3. An ion channel based combinatorial library in 3D subspaces of a 4D DiverseSolutions chemistry space. Channel openers are shown in *black*, blockers in *gray*. The rest of the combinatorial library has been omitted for clarity. Axes are 3D H-hydrogen suppressed BCUTs based on Gasteiger charge (A and D), hydrogen bond donor (A) and hydrogen bond acceptor properties (B). Clockwise from the *upper left* the XYZ axes are ABC, BCD, ACD, and ABD.

clustering of actives in two or three dimensional subspaces of five or six dimensional chemistry spaces shown in Fig. 3 resulted in a novel algorithm for reducing chemistry space dimensionality. Whereas typical methods that reduce dimensionality discard potentially important information, this algorithm identified which axes (metrics) convey information that may be related to the affinity for a given receptor and also identified those axes that appear irrelevant and therefore may be safely discarded. This was done by identifying the axes which tightly group active compounds based upon a cluster-breath normalized value of Chi-squared, computed either from a simple count of actives per bin along each axis or from an activity weighted count of those actives. Multiple binding mode possibilities were addressed by allowing more than one cluster per relevant axis. Receptor relevant subspaces are useful not only for easy graphical visualization of active compounds in the chemistry space, and therefore a visual validation of the chemistry space, but more importantly for the calculation of receptor relevant distances which are essential for identifying near neighbors of actives in focused library design and for comparing libraries.

Stewart et al. have used DiverseSolutions to find a receptor family relevant chemistry space for nuclear hormone receptors that distinguished 907 known NHR ligands from other inactive compounds [39]. Clearly, the next logical step is to use such a space for library design and potentially scaffold-hopping.

In addition to using receptor relevant subspaces as a design method, Pearlman and Smith implemented a list-based nearest neighbor searching algorithm [103] within DiverseSolutions which can also be used for focused combinatorial library design. Currently available versions of DiverseSolutions offer several other novel library design options for focused library design. Ideally suited for target-based library design is a unique cell-based "fill-in" library design option. A set of known active ligands is used to identify "promising cells" in chemistry space. The cells sets of promising cells consist of the ligand occupied cells plus the adjacent surrounding cells up to a user specified cutoff distance (Fig. 4). The chemistry space may have been derived either from a target or target family knowledge database of ligands, the virtual library from which the combinatorial library will be designed or from a standard corporate chemistry space. A reactant-biased product based library design algorithm [24] is then used to design a library, of whatever desired size, which best fills these "promising cells". The degree of target focus is controlled by the number of bins per axis and the number of cell radii from the known ligand is used to define the size of the "promising cell". The focused design approach in DiverseSolutions uses a set of target ligands to score all the compounds in the virtual library based on their distance to the actives and then selects a designed library that optimizes the average virtual activity. An example of using this method to select GPCR

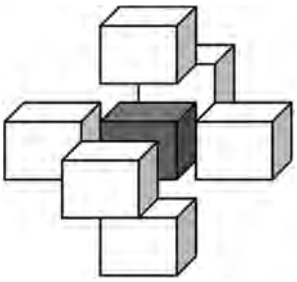


Fig. 4. Representation of "interesting" or "promising" cells. The list of promising cells was chosen to include the cell containing the active ligand plus one layer of the adjacent cells. Additional layers of cells could be added to cover a larger region around the known active to decrease the possibility of missing a hit in the designed library.

compounds for screening to validate the library design approach has been reported by Wang and Saunders [104]. The algorithm also permits use of externally determined activity scores such as those from docking, QSAR models, pharmacophore models or other sources. Additionally, DiverseSolutions offers a novel focused/diverse library design option which yields products that are focused with respect to receptor relevant axes of the chemistry space and are diverse with regard to the receptor irrelevant axes. Use of this algorithm is best limited to individual targets or at best closely related targets since the ligands for an entire family of receptors are unlikely to have the same receptor relevant and irrelevant axes.

The "receptor relevance" of BCUT descriptors has inspired several groups to apply them in conjunction with other methods. Beno and Mason reported the use of simulated annealing to optimize a library design using BCUT chemistry space and four-point pharmacophores concurrently [105]. These authors also have used of chemistry spaces in conjunction with property profiles (unpublished results). The application of such composite methods to target family library design is readily apparent. Pirard and Pickett reported the application of the chemometric method, partial least squares discriminant analysis, with BCUT descriptors to successfully classify ATP site directed kinase inhibitors active against five different protein kinases [106]. Manallack et al. used BCUTS as input parameters to neural networks that selected compounds that targeted specific gene families [107]. Their training sets were derived from known drugs and included three classes: protein kinase inhibitors, GPCR Class A biogenic amines and Class A peptide binding type GPCRs.

2.2. Three-Dimensional Pharmacophore Descriptors

Three-dimensional (3D) pharmacophore descriptors essentially quantify what the medicinal chemist envisions when considering relevant aspects of ligand–receptor binding: several key molecular features/functional groups in a specific relative orientation.

The molecular features encoded in 3D pharmacophore descriptors include hydrogen-bond donors and acceptors, lipophiles, aromatic rings, and acidic and basic moieties. Each of these can play a role in ligand–receptor binding interactions. The relative positioning of combinations of these features within a molecule is determined from 3D conformational models represented by a single low-energy conformation, or by multiple conformations (Fig. 5). Typically, 3D pharmacophores composed of three or four features (3-point and 4-point 3D pharmacophores) separated by three or six distances, respectively, are utilized for computer aided drug design and combinatorial library designs. In order to limit the number of possible 3D pharmacophore descriptors to a manageable quantity, distances are generally binned. Detailed discussions of methods for calculating 3D pharmacophores are found in the literature [34, 74, 108].

If the bioactive conformation of a ligand for a particular receptor is known, then a single 3- or 4-point 3D pharmacophore that is crucial for the binding of that ligand to its receptor may be identified. Other compounds, which contain the 3D pharmacophore of interest, can then be identified via virtual screening or be specifically designed.

Alternatively, bit strings in which the state of each bit (0 or 1) represents the presence or absence of a single 3- or 4-point 3D pharmacophore can be utilized. These 3D pharmacophore "fingerprints" encode all of the 3- or 4-point 3D pharmacophores that can be attributed to a particular molecule (within the limits of conformational sampling resolution). More recently, descriptors that encode weighting factors for individual 3D pharmacophores based on the number of conformations each pharmacophore occurs in divided by the total number of conformations evaluated for a given molecule have been reported to provide enhanced

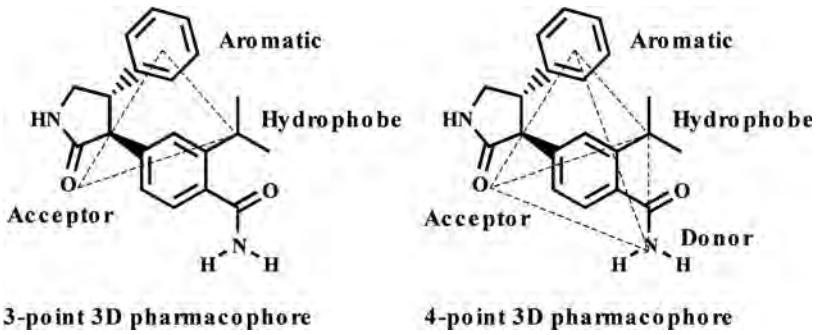


Fig. 5. Schematic illustrating 3- and 4-point 3D pharmacophores. 3-point 3D pharmacophores encode three functional group types and the three distances separating them, and 4-point 3D pharmacophores encode four functional group types and the six distances separating them. Functional group types commonly included are acids, bases, hydrophobes, H-bond acceptors, H-bond donors, and aromatic systems. Distances are assigned to bins (e.g. 2.5–4.0 Å) to limit the individual 3D pharmacophore descriptors to a tractable number, and to aid in comparing the individual 3D pharmacophores.

enrichment in virtual screening exercises compared to simple binary fingerprints for some systems [109]. Schemes that filter out 3D pharmacophores which provide minimal information have also been implemented. Examples of these "insignificant" pharmacophores include those present with similar frequency in both active and inactive compounds relative to the same biological target, and pharmacophores that contain only lipophilic features [109]. Methods for evaluating the potential information content of individual 3D pharmacophores for have also been employed [110, 111]. Pharmacophore fingerprints can be generated for sets of compounds by performing a logical OR operation on the fingerprints for the individual molecules. Furthermore, the similarity of one molecule to another can be assessed by calculating the Tanimoto [34, 112, 113] coefficient of the 3D pharmacophore fingerprints for the two molecules. Additionally, pharmacophore fingerprints can be utilized as descriptors for QSAR model development. When individual pharmacophore descriptors (i.e. specific bits in a pharmacophore fingerprint) are shown to be statistically significant in a QSAR model, the corresponding structural features in the compounds used to derive the model can often be readily identified and designed into or out of subsequently synthesized molecules.

Pharmacophore descriptors have been utilized to design both diverse compound libraries and libraries focused on specific biological targets. Target family library design schemes employing 3D pharmacophore descriptors have also been reported. Selected examples demonstrating the application of 3D pharmacophore descriptors to target-focused and target family library design projects are described below.

2.2.1. Target-Focused Library Design with 3D Pharmacophore Descriptors

Several early literature reports highlighted the effectiveness of 3D pharmacophore descriptors for similarity searching [34, 114] and predictive QSAR model development [74], and suggested the possibility of their use in the context of target-focused library design. Among these is a study reported by Pickett and co-workers [114] which utilized 3D pharmacophores derived from the capped RGD tripeptide to identify a set of compounds enriched in known fibrinogen receptor antagonists. Generation of the set was determined from the number of 3D pharmacophores in common between the reference compound and each database compound. The authors noted that a similar methodology could be applied to combinatorial library design. Mason et al. reported a "Tanimoto-like" coefficient for calculating intermolecular similarity based on 3- and 4-point 3D pharmacophore descriptors [34]. Work by McGregor and Muskal utilized PharmPrint™ 3D pharmacophore descriptors (3-point 3D pharmacophores) and partial least squares with principal component analysis to develop models for predicting estrogen receptor activity, and also for selecting "drug-like"

screening libraries [45, 74]. These seminal studies provided a foundation for the efforts described below.

The following example of a target-focused library design demonstrates 3D pharmacophore methodology in which fingerprints from both active compounds and known *inactives* are utilized to identify which 3D pharmacophores are important for ligand/receptor binding.

Utilizing a set of 43 known α1-adrenergic receptor ligands with Ki values <5 nM (actives), and a set of 62 compounds with Ki values > 5 μM against α1 receptor subtypes (inactives), Bradley and coworkers derived a 3D pharmacophore ensemble model that correctly identified 80% of α1 actives and only 10% of the inactives as active compounds in validation studies [110]. This ensemble model was composed of a set of five hundred 2-, 3-, and 4-point 3D pharmacophores with the highest "information content" found in the analysis of the active and inactive compounds in the model training set. The information content for each 3D pharmacophore was calculated using an equation derived from information theory. Essentially, those 3D pharmacophores which occur with high frequency in known actives, but are absent, or occur with low frequency in known inactives for a particular target are high in "information content" and may be used in combination to discriminate actives from inactives.

Bradley and coworkers used the 3D pharmacophore ensemble model to filter a virtual combinatorial library of 3,924 N-substituted glycine peptoids [115] containing three known α1 actives down to a set of 639 products. Using a "cut-down" technique, a 160-compound combinatorial library was designed in which the number of compounds that passed the ensemble model filter was maximized. This library contained two of the three known actives present in the original 3,924-compound virtual library. This represents a substantial enrichment ((2 actives/160 products) × 100 = 1.25% vs. (3 actives/3,924 products) × 100 = 0.076%).

Beno and Mason reported an alternative design approach based on 3D pharmacophore frequency counts that used the same set of 43 known α1 ligands, and a virtual library of 10,648 N-substituted glycine peptoids [75, 115]. The virtual library contained at least three products known to be active at α1. In this case a library of 343 products (7 R1 × 7 R2 × 7 R3) was selected with a simulated annealing procedure that maximized the similarity of the normalized 4-point 3D pharmacophore frequency distributions of the active α1 ligands and the products comprising the selected virtual library subset. In this case, one of the three known α1 actives was found in the final library. Comparison of these results to those of Bradley et al. [110] suggests the importance of including inactive compounds when deriving 3D pharmacophore-based computational models. The inclusion

of 2- and 3- as well as 4-point 3D pharmacophores may also lead to models with improved discriminating power.

In a retrospective study utilizing ensembles of 3D pharmacophores selected based on their ability to discriminate active from CDK2 antagonists from inactive compounds, Bradley et al. [116] highlighted the enhanced effectiveness of informative screening library design strategies compared to schemes utilizing diversity/similarity. In each round of model refinement, the enrichment achieved through application of pharmacophore descriptors and informative design techniques was at least 1.6 and as much as 4.9 times that obtained with diversity–similarity methods. Although the compound selections in this example were not subject to combinatorial constraints, the strategy employed is equally applicable to combinatorial library design.

An example of target-focused library design utilizing QSAR models derived from 3D pharmacophore descriptors was recently reported by Deanda et al. [117]. Three QSAR models predictive of pIC_{50} values against ATK1, Aurora-A, and ROCK1 kinases were generated utilizing PharmPrint™ 3D pharmacophore methodology as modified by Deanda and Stewart [118]. These models contained between 1,800–3,200 compounds in their respective training sets, and had non-cross-validated r^2 values ranging from 0.66 to 0.81 with standard errors of prediction in the 0.51–0.56 log unit range. Two virtual libraries, each based on multiple hinge-binding scaffolds, and containing approximately 43,000 and 97,000 virtual products, respectively, were enumerated and scored against the three PharmPrint™ QSAR models. Based on the distributions of predicted activity for each large virtual library against each of the three kinases, small Aurora-A and ROCK1-targeted libraries were designed to maximize the predicted average pIC50s of each selected subset of compounds against one of the two kinase targets. From the Aurora-A focused library, 43 compounds were ultimately synthesized and tested. Of these, 42 were predicted to be active by the Aurora-A QSAR model based on a predefined threshold for activity of $pIC_{50} \geq 6$. From the set of 42 compounds, 18 had actual measured pIC_{50} values ≥ 6. This compares favorably to results for an Aurora-A library of 68 compounds synthesized as a control. None of those compounds met the criterion for activity when tested. For the ROCK1-focused library of 297 compounds, 144 were predicted to be active by the ROCK1 QSAR model, and 44 were actually active based on the same threshold for activity that was used for the Aurora-A compounds. In a ROCK1 control library of 749 compounds, 38 compounds were active. For each kinase target-focused library, significant enrichment in active compounds relative to "hand-selected" control libraries was achieved through utilization of PharmPrint™ 3D pharmacophore-based QSAR models.

Multiple-point 3D pharmacophore descriptors are also useful for designing libraries when crystal structures are available for the target(s) of interest. Fingerprints consisting of 3D pharmacophores that are complementary to binding site features can be created and used in conjunction with docking studies to select products with optimal shape and pharmacophoric features from virtual combinatorial libraries [105, 119, 120]. This technique has been used to design Factor Xa [119] and cyclin-dependent kinase (CDK-2) [120] focused libraries. More recently, two new methods for creating 3D pharmacophore descriptors for protein active sites have been reported. SitePrint [121] derives multiple-point 3D pharmacophores for protein active sites from clusters of docked small-molecule probes, and FLAP [122] uses points obtained from GRID molecular interaction fields to generate active site and ligand 3D pharmacophores. Each of these has potential utility for both target-focused and target-class library design. These methods could be extended to multiple targets within a target family by calculating multiple-point 3D pharmacophore fingerprints that are complementary to the binding/active site for each target, and then performing a logical AND operation on all of the fingerprints to determine their intersection. The resultant fingerprint of receptor-complementary, common 3D pharmacophores could then be used in conjunction with a docking algorithm to select products from a virtual combinatorial library.

2.2.2. Target Family Library Design with 3D Pharmacophore Descriptors

The examples provided in the previous section all use (or could use) multiple-point 3D pharmacophores from sets of known active ligands, and in some cases, inactive ligands as well, to generate models for library design. These models discern, to varying degrees, the 3D pharmacophores that are relevant to ligand/receptor binding, and may be derived from large numbers of diverse compounds covering many diverse chemotypes active against different, but related targets (e.g. GPCRs). These models represent the common ligand features that may be recognized by different members of a biological target family. Thus, they are well suited to the design of target family libraries, which emphasize the commonalities between related targets.

Multiple-point 3D pharmacophore fingerprints can also be used to calculate the similarity between pairs of molecules using the Tanimoto coefficient, or similar metrics. Individual target-focused libraries may be designed by maximizing the 3D pharmacophore similarity of product compounds to ligands known to be active against the target of interest [114]. Target family combinatorial libraries may then be created by combining smaller libraries focused to individual, related targets.

Several literature reports [34, 35, 123] and at least one patent application [124] describing target-class library design strategies

which employ 3D pharmacophores have been published, and a selection of these is discussed below.

Mason and coworkers reported one of the first examples of a target family library design utilizing 3D pharmacophore descriptors [34]. In this example, the authors designed a set of GPCR-targeted libraries based on Ugi chemistry [125]. A key feature of the design is the incorporation of a GPCR privileged substructure in each of the combinatorial products. GPCR privileged substructures are chemical moieties that occur with high frequency in the ligands of multiple GPCRs [48]. Examples include biphenyl tetrazole, indole, and biphenyl-methyl groups. Methods to derive them will be discussed below.

In the published example, 502 compounds from the MDL Drug Data Report (MDDR) [126] which were active against a GPCR target and also contained a biphenyl tetrazole moiety were used to generated a "privileged" 4-point 3D pharmacophore fingerprint. A privileged 4-point 3D pharmacophore is a 4-point 3D pharmacophore in which one of the four molecular features is a privileged moiety, and the other three are members of the standard set of feature types (H-bond donors, acids, etc.) In this case, the privileged feature was represented by the centroid of the biphenyl tetrazole moiety in each compound. The 4-point 3D pharmacophore fingerprint for the set of 502 GPCR ligands was the union of the fingerprints of the individual molecules, and represented ~161,000 privileged 4-point 3D pharmacophores.

Utilizing a simple greedy algorithm, a set of 22 acid reagents (along with 12 aldehydes and 8 isonitriles to yield 2,112 products) was selected to maximize the intersection of the privileged 4-point 3D pharmacophore fingerprint of the combined combinatorial products with the GPCR privileged 3D pharmacophore fingerprint derived from the MDDR compounds. Approximately 49% of the GPCR privileged 3D pharmacophores found in the GPCR privileged fingerprint were covered by the products of the optimized reagents. Subsequent libraries were designed to cover the 4-point 3D pharmacophores present in the GPCR reference fingerprint that were not covered by the original library.

The GPCR target family bias in this first example was achieved using two key design elements. The first is the incorporation of a GPCR privileged substructure, and the second is maximal coverage of GPCR privileged 4-point 3D pharmacophores found in known GPCR ligands. This approach weights all of the 4-point 3D pharmacophores found in the set of known GPCR ligands equally in terms of their importance for receptor/ligand binding.

In a similar target family design effort utilizing a set of 3,321 GPCR ligands with reported in vivo activity, Lamb and coworkers identified a set of 1.8 million 2-, 3-, and 4-point pharmacophores predicted to be important for GPCR binding [123]. However, in

this case, rather than treating all of the 3D pharmacophores found in the known actives as equally important, only those 3D pharmacophores that occurred in at least ten active compounds were included in the GPCR reference 3D pharmacophore key. Utilizing this reference key, a library composed of 7,865 products that covered 66% of the 1.8 million GPCR-relevant 3D pharmacophores was designed.

This methodology was extended to the design of a library of approximately 14,000 compounds [35]. In this case, a filtered collection of 2,785 GPCR ligands reported in the MDDR was used to derive a set of 1.1 million 3- and 4-point pharmacophores that were present in at least ten of the MDDR GPCR compounds. Each pharmacophore in this GPCR "pharmacophore-space" was required to be present in at least ten library compounds to be considered "covered" by the library. Following initial synthetic efforts, an iterative "build-up" approach was utilized in which additional scaffolds were evaluated for their potential to provide coverage of pharmacophores not already included at the required frequency in the library. For the most promising scaffolds, monomer selections were made based on the ability of the resultant compounds to maximize coverage of the missing or under-represented pharmacophores. The library of ~14,000 compounds was screened against the μ-opioid receptor resulting in a 2.6% hit rate. In this example, and the preceding one, use of a 3D pharmacophore "frequency count" in the analysis of known active ligands improves the odds of identifying 3D pharmacophores that are actually relevant to ligand/receptor binding [127].

Target family combinatorial libraries are intended to contain products that are active against multiple members of a family of biological targets. However, activity/potency is not the only concern. Optimally, the combinatorial products should be as "drug-like" as possible, with minimal ADMET liabilities. This is a difficult goal to achieve. However, computational models utilizing multiple-point 3D pharmacophore descriptors may be used to address some of these issues as well. A 3D pharmacophore model for PGP substrates has been published [128], key pharmacophores common to many CYP3A4 inhibitors [129, 130] have been identified, and a 3D-QSAR model utilizing 3D-pharmacophore fingerprints to predict CYP2D6 metabolic stability has been reported [131]. Specific 3D pharmacophores are also key components of a QSAR model predictive of hERG inhibition [132]. These may be utilized for target family library design. For example, one might design a GPCR target family library in which the coverage of GPCR-relevant 3D pharmacophores is maximized, while the coverage of PGP substrate pharmacophores is minimized in the selected products.

There are several reported examples of target-focused and target family library design work that utilize multiple-point 3D

pharmacophore descriptors, and recent reports suggest that interest in this technology is unabated [133, 134]. Multiple-point 3D pharmacophores are well suited for these design efforts due to their ability to encode common features recognized by receptors/enzymes and displayed by small molecules. They can be utilized with protein structures, collections of ligands, or single ligands. Models that predict desirable biological activity or potential ADMET liabilities can be developed with 3D pharmacophore descriptors. As interest in target-focused and target family libraries continues, multiple-point 3D pharmacophore descriptors should see extensive use in combinatorial library design.

2.2.3. Privileged Substructures for Receptor Target Family and Receptor Focused Design

Since creation of combinatorial libraries directed towards families of receptors such as GPCRs, kinases, nuclear hormone receptors, proteases, etc., has largely replaced the generation of diverse libraries, it is worth examining one of the specific design methods for target family libraries in detail. The original concept of "privileged substructures" was put forth by Evans [32] and reviewed by Patchett [33]. They described privileged substructures as those substructures found in ligands across a set of diverse receptors. Further elaboration of the privileged substructure was postulated to lead to selectivity toward a specific target receptor. While Evans and Patchett evolved this concept within a relatively narrow class of GPCR ligands, subsequent literature methods for finding privileged substructures and common practice in combinatorial chemistry library design not only expanded on their original analysis, but also modified the definition of "privileged structure" to that of commonly occurring fragments within ligands associated with a target receptor family.

Clearly, the term "privileged substructure" has taken on a meaning beyond Evans original intent. It has become identified with those substructures found to be promiscuous within a given target family and carries the implication that these substructures are specific to that target family. If such substructures exist, off-target affinities might be avoided early in the discovery process and thereby avoid complications as promising compounds are developed into drugs. If these substructures can be identified, they potentially provide cleaner starting points than do the more promiscuous structures.

Various methods have been employed to find so-called target family privileged substructures that are postulated to be selective for a given target family, but promiscuous within that same family of targets. The majority of the commonly used methods are ligand based and include frameworks analysis [135], 4-pt pharmacophores [136], ClassPharmer^{TM} [137] substructure class generation [138, 139] and Pipeline Pilot [140] substructure generation using ISIS-like keys [141, 142]. These fragments have been used not only for target family combinatorial library design [35,

36, 143], but also for virtual screening [144] and for focused screening deck design [145].

Typically, to find privileged substructures, one first selects a set of literature structures such as "Family A" or "Family B" GPCR non-peptide ligands in the MDDR [146] and performs frameworks analysis. One then performs maximum common substructure (MCS) analysis on all the frameworks and removes the ligands with that substructure from the superset of ligands. The frameworks analysis, MCS analysis and ligand removal are iterated until 90% of the ligands are have been accounted for. For a ligand set from the 1999 version of the MDDR [126], 15 SLNs (SYBYL line notation structures) [147] accounted for 90% of the Family A and B GPCR ligands [148].

Typically these target family privileged structure analyses have attempted to find minimal ligand substructures that have frequent occurrence within a particular target family. However, this can very easily lead one away from truly privileged substructures and toward those which are merely "drug like" and/or promiscuous protein binders. Consider the comparison of the often cited [34, 135, 138, 149], GPCR privileged substructure, biphenyl, and its analog 2-tetrazolobiphenyl (Fig. 6). A substructure search [138] of the 2004 version of the MDDR found that 2-tetrazolobiphenyl appeared in 1,046 compounds, all of which fell into the activity classes related to the Angiotensin II receptors. The biphenyl substructure was found in 5,658 compounds spanning 311 activity classes which included not only a significant number of GPCRs, but also a host of other targets. Although biphenyl may be classified as privileged due to its frequent appearance in GPCRs, it is clear that true privilege does not arise until the tetrazole moiety is included. Biphenyl itself is likely only to be a privileged protein binding element.

While some of the literature studies involving target family privileged substructures compare the fragment occurrence frequency of GPCR privileged substructures with non-GPCRs as a whole among known drugs [37], little analysis has been done on the selectivity of these substructures with respect to other target families [138, 143]. In part, this has been due to the difficulty of

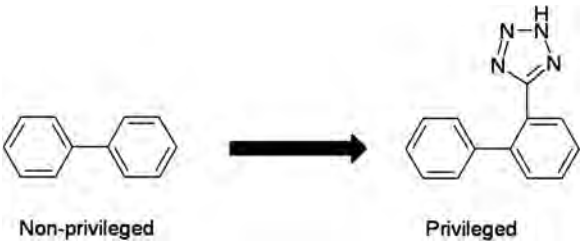


Fig. 6. An example of a non-privileged versus a privileged substructure: biphenyl and its analog 2-tetrazolobiphenyl.

collecting or extracting the target family ligand sets from commercial drug databases and corporate collections. Recently however, the publication of an ontology of pharmaceutical ligands by Schuffenhauer [41] removed this roadblock by mapping MDDR activities to published target ontologies. Additionally, the need for target family knowledge databases to drive target family based library design has resulted in a number of commercial target family databases from companies such as Aureus [42], Jubilant [43], and Sertanty [44].

Schnur and co-workers [138] demonstrated that care must be taken in target family library design to ensure that the privileged substructures chosen are truly selective for the target family of interest. In that study, ClassPharmer™ was used to generate substructures for MDDR derived ligand sets of Class A GPCRs, nuclear hormone receptors, kinases, ion channels, and serine proteases. The compound sets for each receptor family were filtered through the substructure sets for each other receptor family. Both the percentages of occupied substructures for each cross-filtering and the percentages of compounds from each set that occupied other target family substructure sets were examined. Examples of GPCR "privileged substructures" that were not selective for just the GPCR receptor family were provided.

Nonetheless, privileged substructures, if used as Evans originally defined them, [32] can provide useful starting points for library design [150]. Their use in designing GPCR libraries [36, 151] has been reviewed and new examples continue to be published [152]. Undoubtedly the design literature will continue to be enriched by further examples using not only GPCRs, but other target families as well. As Evans pointed out, it is the elaboration of the privileged substructure that provides target selectivity [32] and elaboration of a substructure is in fact the essence of library design. Whether a library can be designed that is selective for a specific receptor family is a subject for ongoing debate, but target family libraries certainly enhance corporate collections and are frequently assayed against other receptor families than those for which they were intended.

2.3. Structure Based Methods

Structure based drug design (SBDD) has been an integral tool in the drug discovery process for more than 20 years and, with the increasing availability of experimental protein structures, its influence continues to grow. Several recent reviews [153–156] have highlighted the impact of structure on the design of compounds targeted toward the kinase, aspartyl protease, metalloprotease, and nuclear receptor families of proteins. Incorporation of structure into the design process has the powerful effect of focusing the chemical space around a particular scaffold into a space most relevant for the target of interest. However, inherent to the successful application of SBDD is the generation of multiple variants

around a designed target molecule and iteration based on the results of each generation of design. Inaccuracies in the docking tools and scoring functions preclude the selection of a single, optimal molecule and require a certain amount of chemical serendipity to insure successful outcome.

Structure-based combinatorial library design attempts to incorporate structural focusing into the library design process. At the most basic level, this can be done simply by fully enumerating a virtual library and docking those products in a serial manner. However, this strategy is only efficient for small libraries because the time required to dock the library grows as a multiple of the number of molecules. To address this, several methods (Fig. 7) have been developed that eliminate the need for a fully enumerated library. Most of these methods process the core and each of its potential substituents independently thus reducing the timing to be merely additive with the number of products rather than multiplicative.

One of the earliest methods to take this approach was CombiDOCK [157]. CombiDOCK is a workflow based on the DOCK [158] program which begins with the docking of unsubstituted cores into the protein binding site. Once a core is oriented, substituents are grown individually from each R-group attachment point on the core and scored within the receptor. The best scoring substituents for each attachment point are then combined, checked for intramolecular clashes, and retained if no clashes are found. The method was validated retrospectively with a 1,000– compound hydroxyethylamine library which was targeted toward Cathepsin D, and showed fourfold enrichment relative to random selection. The method has also been applied to the design of a selective nuclear hormone receptor [159] and inosine

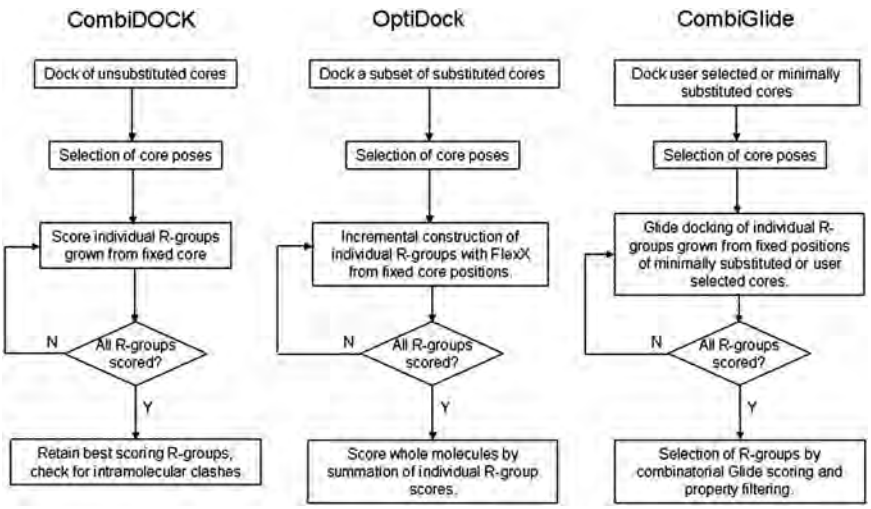


Fig. 7. Typical workflows for the combinatorial docking programs CombiDOCK, OptiDOCK, and CombiGLIDE.

monophosphate [160] inhibitors. One of the primary limitations of CombiDOCK is the initial placement of the unsubstituted core. Docking the bare core molecule can lead to unrealistic binding modes because interactions and clashes with its substituents are not present. Because the core is held fixed in the subsequent evaluation of elaborated molecules, any error in core placement will be propagated and affect compound scoring. More modern algorithms, like OptiDOCK [161] and CombiGLIDE [162–165], attempt to address this issue using different sampling methods.

The OptiDOCK program from Tripos [147] samples potential core placements by first docking a small subset of fully enumerated compounds and retaining diverse and representative core placements. The core is then fixed in place and the incremental construction protocol FlexX [166–168] is used to build and score substituents sequentially for each R-group position. A whole molecule score is then derived by summing the scores for all R-groups on the molecule. A related method FlexX^{C} [169–171] that uses the same incremental construction has also been reported, but instead of starting with a defined core, it begins with a base fragment selected from a substituent list. While this approach is more efficient than serial FlexX [172, 173] since the base fragment placements can be stored and reused, it can lead to inconsistent core placement because the remainder of the molecule is incrementally constructed from the base fragment. OptiDOCK was shown to accurately reproduce binding modes produced by FlexX serial docking. When validated against a library thrombin targeted library the OptiDOCK energies were found to correlate with pIC50 with an r^2 of 0.58 [161]. With the exception of a solitary application to kinase library design [174], no subsequent applications of this methodology have been reported.

CombiGLIDE, recently developed by Schrodinger [87], follows a similar core placement plus side chain scoring procedure. Core placement can be derived from its crystallographic position or sampling for all potential placements within the binding site. As with CombiDOCK and OptiDOCK, each potential R-group position is sampled individually with fully substituted compounds selected via a proprietary selection algorithm. In addition to selection by docking score, the user also has the option to filter compounds by calculated ADME properties to potentially focus the library into a more drug-like space. CombiGLIDE has been demonstrated to select known active compounds in retrospective studies against the estrogen receptor [164] and p38 [175]. No prospective design studies have yet been reported.

Surveys of the literature around the receptor based methods mentioned above and other related methods (reviewed recently by Zhou [176]) make it is apparent that they are sparingly cited. To some extent, the shallow penetration for many of the early programs into the drug discovery process is a reflection of the

difficulty non-experts have experienced in setting up the programs and getting reasonable results. Much of this has been resolved through more modern software engineering and interface design like that of OptiDOCK and CombiGLIDE. Of more concern for both the newer and older programs, is a lack of references citing prospective designs that elucidate active compounds in a prospective manner, bringing into question the general validity of the approach. There have been few reports detailing the use of these methods in an iterative fashion, i.e. using the data from one generation to focus the design of the next. Nonetheless, while these tools may not be specifically mentioned, the general strategy of incorporating structure and iteration into library design has been exemplified in the recent literature.

The discovery of novel EphB4 kinase inhibitors using the ALTA (anchor-based library tailoring) approach has been reported by Kolb et al. [177]. The ALTA procedure begins with fragment decomposition of a pool of compounds by cutting rotatable bonds. The resulting fragments are then filtered and docked into the protein active site. The top scoring fragments serve as probes to screen the original pool of molecules using graph-subgraph isomorphism. Compounds that pass the screen are flexibly docked using the fragment as an anchor, minimized and ranked using a linear interaction energy with continuum electrostatics model. Screening of a 782,202 compound library against EphB4 kinase resulted in the selection of a 21,418 compound subset. Forty compounds were tested in a FRET-based assay; ten interfered with the assay, and of the remaining thirty, six had IC_{50} values $< 100\ \mu M$.

A two phase design strategy resulted in the discovery of submicromolar inhibitors against the β-hydroxyacyl-acyl carrier protein dehydratase of *Heliobacter pylori* (*Hp*FabZ) [178]. In the first phase, 39 analogs of two previously identified leads were designed using crystal structures of the lead compounds in *Hp*FabZ. The biochemical potency of one lead series was increased 4–28-fold during this design phase although no improvement was observed for the second. Crystal structures for six analogs in the active series were determined and formed the structural basis for focused library design in phase two. A 280 compound virtual library was evaluated using the program LD1.0 [179], which scores based on docking score, structural diversity and drug-likeness. Twelve compounds were selected for synthesis. The library showed a 67% hit rate with the most potent analog having a 46-fold improvement in binding affinity relative to the initial lead.

The protein-protein interface between p53 and HDM2 has been probed via docking of weak hits to the HDM2 binding site followed by small library design and synthesis [180]. A set of six isoindolinone hits that ranged in potency from 27 to 92 micromolar were docked into HDM2 to yield a large number of

potential binding modes. Twenty-four unique core binding modes were retained and used to enumerate a series of R1 and R2 substituents derived from commercially available reagents. The virtual compounds were scored using Skelgen [181] and filtered based on their ability to form at least one hydrogen bond. Fifty-seven compounds were selected for synthesis, of which 43 were evaluated. Thirty-eight of these had IC_{50}s < 500 μM with the most potent having an IC_{50} of 14 μM. Based on the results from the first library, a second combinatorial library was prepared and led to an inhibitor with and IC_{50} of 5.6 μM.

2.3.1. Structure Based Methods Employing Fragment Based Design

Historically, structure based library design focused on lead optimization of compounds already drug like in size thereby making further optimization difficult. An alternative strategy shown in Fig. 8 begins with the identification of weakly binding small fragments that serve as the foundation for larger, more potent binders. This process is commonly referred to as fragment based drug design (FBDD) [182]. Although the initial lead fragments possess low absolute binding affinity, the amount of binding energy per atom, or ligand efficiency, is often comparable to that of highly potent, but much larger lead like molecules [183]. If one can retain high ligand efficiency as the small fragment is elaborated, potency will increase as additional interactions are made with the protein target. However, the successful prosecution of a FBDD campaign relies on several key components: a suitable library of fragments to screen, sensitive methods to detect binding, and structural information to guide the design of compounds based on the active fragments. Once these components are in place, FBDD has been demonstrated as a viable lead finding

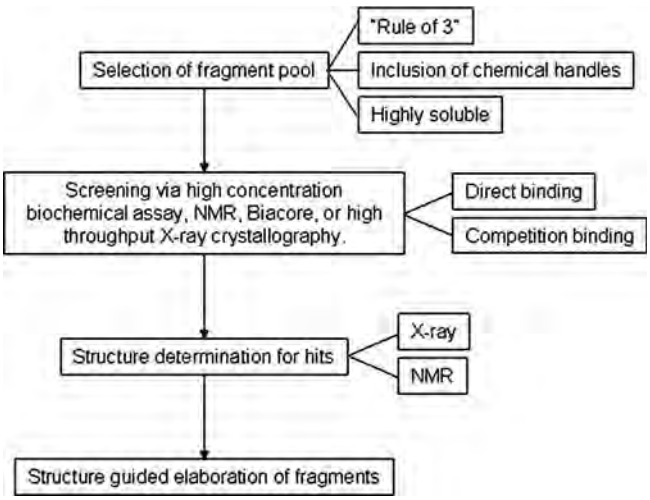


Fig. 8. Schematic representation of the fragment based drug discovery (FBDD) process.

strategy for a diverse set of targets including: FK506 binding protein (FKBP), B-cell CLL/lymphoma 2 (Bcl-2)/BCL-2 like 1 (Bcl-xL), Heat shock protein-90 (HSP90), Aurora kinase, β-site APP cleaving protein (BACE), and leukocyte function associated antigen-1 (LFA-1) [184].

The composition of a screening library targeted toward FBDD is markedly different than one designed around lead like compounds. The goal is to find starting fragments that occupy high affinity regions within the binding site for subsequent elaboration rather than modification of a larger lead structure that already occupies much of the site. As a result, criteria generally applied during compound filtering, like Lipinski's "Rule of Five" [185] and variants there-of, are not directly applicable to fragment selection because they were derived from larger drug-like molecules. Instead, Congreve et al. [186] proposed a "Rule of Three" based on analysis of fragment hits identified by screening with X-ray crystallography. They found that, on average, fragments should have a molecular weight < 300, hydrogen bond donor count < 3, hydrogen bond acceptor count < 3 and a cLogP < 3. Their data also suggested that a rotatable bond count < 3 and a polar surface area < 60 may also be useful. These selection criteria have subsequently been validated in a study by Hajduk [183], where a set of 18 optimized inhibitors were deconstructed and the activities of the component fragments evaluated. The discovery of the Bcl-xL inhibitor ABT-737 was presented as an example system. A striking correlation was found between the molecular weight of each fragment and the K_d against Bcl-xL. The binding efficiency index, which is calculated for each compound by dividing the pK_d by molecular weight, was nearly constant from the smallest fragment to the optimized compound and spanned a molecular weight range of 293–813. This suggested a simple linear relationship between binding affinity and the number of heavy atoms with each atom contributing approximately 0.3 kcal/mol. The same trend was apparent for the other 15 targets used in this analysis and suggested that the minimum size range for efficiently binding fragments to be observable biochemically or biophysically should be about 10–20 atoms, which is consistent with the "Rule of Three".

In addition to filtering based on the "Rule of Three", the selection of fragment compounds that possess a functional group amenable to library synthesis can greatly facilitate the construction of compounds for lead follow-up. Solubility is also an important consideration because these fragments will need to be screened at high concentration. Even with these constraints, a large number of potential molecules are still available, thus necessitating subset selection either by clustering, diversity selection, scaffold-based classification [187], or functional group classification. This can be accomplished using tools already well established for library design.

Once a library of starting fragments has been defined, an assay must be designed that can detect binding well outside the range usually considered during lead optimization, because these fragments typically have binding affinities in the high micromolar to millimolar range. A number of strategies, which include high concentration screening, NMR, surface plasmon resonance (SPR), and high throughput X-ray crystallography, have been employed to screen these fragments and have been reviewed by Siegal [188]. The seminal work in this field was done by the Abbott group who utilized their SAR by NMR strategy in search of FKBP inhibitors [189]. In this early example, the protein was isotopically labeled and the compounds screened as mixtures. Compounds were identified with micromolar affinities and subsequently transformed into nanomolar inhibitors by linking together two smaller fragments. Since NMR is quite sensitive to chemical perturbation by the fragments, but does not give a high false positive rate, this patented method remains a very viable solution. Also, if the protein NMR is assigned, 3D information about ligand binding can be discerned from the experiment. The method does, however, have some significant limitations: the requirement for labeled protein is high, the protein must be less than about 40 kDa and it must be soluble in the presence of high fragment concentration. To address these limitations, several NMR methods have been developed that look at changes in the ligand spectra rather than the protein. The saturation transfer difference (STD) [190] and WATERLOGSY [191] methods are the most commonly utilized and rely on the transfer of magnetization from the protein to the ligand. Although there are fewer protein limitations with these methods, it must still be soluble in the presence of ligands and be >20 kDa for good sensitivity. A related method, target immobilized NMR screening (TINS) [192], screens an entire library across a single immobilized protein sample, thereby potentially addressing both the issues of protein availability and stability. In addition to assessing direct binding, these methods can be run in a competition binding mode by including a chase ligand known to bind in the active site. Those compounds that do not bind in the presence of the chase ligand are likely to be competitive with it. This allows the differentiation of fragments targeting a known site from those that bind to an alternate site.

NMR has been used most frequently as a screening tool, but surface plasmon resonance (SPR) has also begun to show utility in the fragment space. Advances to SPR instrumentation have enhanced the sensitivity of the technique to the point where fragments can now be detected [193, 194]. SPR measures the mass change at the liquid-solid interface by flowing one binding partner across a second that is immobilized on a surface and then detecting a change in refractive index at the surface.

This detection method allows substantial flexibility in assay format because one can immobilize the protein, a ligand to be competed, or the compounds themselves, thus making both direct binding and competition binding studies possible. The typical screening protocol fixes the protein to the chip which is then exposed to soluble ligands [195], but new surfaces have recently become available which contain arrays of immobilized ligands that are assayed for their ability to bind to a soluble protein. For example, Graffinity [196] has developed instrumentation and chemistry to create and screen chemical microarrays of up to 10,000 small molecules.

SPR has evolved into a relatively high throughput screening process, but it does not provide any detailed structural information. NMR can be used to glean this type of data if the protein is appropriately labeled and assigned, but it is low throughput. In order to efficiently gather structural information, many have turned to high throughput X-ray crystallography. In fact, this technology is the basis for companies like Astex Therapeutics, Plexxikon, and SGX Pharmaceuticals. High throughput X-ray crystallography can be used as the primary screen, but the technical issues involved in setting up the required number of crystal trials limit the manageable size of the screening library to <1,000 [197]. Instead, the fragment library is usually screened first by a higher throughput method with crystallization attempts made only on the confirmed hits.

Once the hits have been identified and some structural information is in hand, the challenge then falls to both the computational and organic chemists to decorate these fragments to provide additional productive interactions with the protein while maintaining good ligand efficiency. This is necessarily an iterative process, but brings to bear many of the structure based library design techniques described above. Several clear examples of successful applications of this strategy have been described in the recent literature and reviewed [198]. Many targets are approached by simply treating the fragment as the initial anchor point and then designing multiple 1D libraries based on it, with subsequent libraries building on the results of the first. This is easily accomplished with structure based design tools like CombiDOCK, CombiGLIDE, FlexX, etc. However, there are several examples of the linking of two small fragments in binding distinct subsites, leading to high affinity lead compounds. For example, the design of Abbott's Bcl-xL pre-clinical candidate ABT-737 (Fig. 9), began from two disconnected fragments with Kds of 0.3 mM and 4.3 mM that were shown to occupy proximal sites on the protein surface. Several rounds of library design and medicinal chemistry optimization ultimately identified an appropriate linker and more tightly binding fragments that led to ABT-737 with a Bcl-xL Kd of 36 nM [199]. This example, along with the others cited

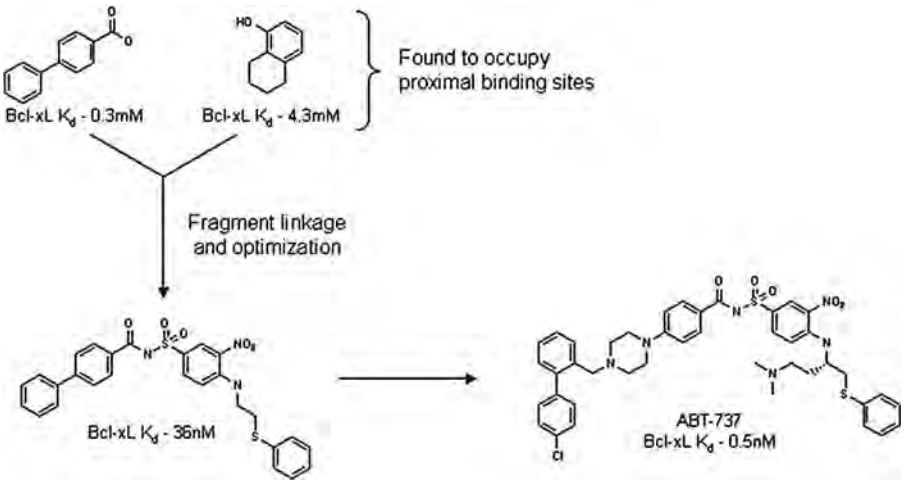


Fig. 9. An example of a successful application of FBDD toward the discovery of the Abbott Bcl-xL antagonist ABT-737. The *topmost* fragments were discovered by NMR screening and found to bind in proximal sites. Linking these fragments and subsequent optimization of the led to a 36 nM inhibitor shown in the *lower left*. Further optimization ultimately gave rise to the pre-clinical candidate ABT-737.

in Law [198] and Hadjuk [184], demonstrate that with the necessary tools and commitment, fragment based design is a viable alternative to high throughput screening. It should be noted, however, that there needs to be a clear path forward from the initial identification of hits to structure determination and compound synthesis based on these structural results. Fragment based drug design requires multiple rounds of optimization before a suitable path toward robust affinity is identified and one must resist the temptation to drop a fragment because of negative early results. A serious commitment by the biophysical and chemistry teams involved in the project is required for a fragment based campaign to progress.

2.4. Comparative Library Diversity, Scaffold Design, and Bio-isosteres

While interest in compound diversity as a means of designing large libraries has faded, the recent literature still contains reports of methods that address the comparative diversity of libraries and of collections. The need for securing intellectual property via compound novelty has become almost as important a criterion for combinatorial library design as pharmaceutical relevance. Originally, Pearlman's DiverseSolutions was designed specifically with tasks such as library or collection comparison in mind, but Pearlman focused on property based comparisons using BCUT descriptors. The current trend is to actually analyze diversity space coverage of reported combinatorial libraries in terms of chemotypes or scaffolds. The annual collections, published by Ron Dolle, of combinatorial libraries that have appeared in the literature since 1997 has facilitated this effort [200–210]. Recent publications in the field of library design have thus focused on

methods for finding bio-isosteric replacements for known chemotypes, molecule cores or scaffolds. Replacement of side chains by bio-isosteres for improvement of ADMET profiles for drug candidates is also an area of ongoing interest.

Geysen and colleagues [211] developed a method for comparing libraries that they described as enabling "soft scaffold hopping". Most 2D similarity methods such as UNITY [147] fingerprints, MDL MACCS keys [141] or DiverseSolutions BCUTs that are commonly used for library comparisons are based on properties which arise solely from a molecule's connectivity table. Such methods do not adequately account for global properties such as molecular size and shape; however 3D methods such as shape matching [212] and pharmacophore searching [34] require knowledge of the bio-active conformation. If this is unknown, then the 3D methods are less effective. All of these methods make comparisons via a structure by structure approach. By contrast, Geysen et al. proposed a library level approach. Their similarity metric encompasses 3D information specific to each molecule's accessible conformations by focusing on the spatial orientation of the R-groups provided by the common scaffold/s found in the library. This spatial orientation of the R-group along a contact surface of a protein is the key factor for the binding interactions. Provided that the scaffold or template core does not also interact with the receptor, the metric is appropriate for scaffold hopping. For their analyses, Geysen et al. defined a diversity triangle based on the distances between the diversity points of a scaffold having attached three R-groups. The lengths of the triangle were then translated into x, y, z coordinates that designate a point in diversity space. Using a cell based approach, the degree of similarity between two libraries was assessed on the basis of the number of diversity cells the libraries had in common and was given as a percent overlap. While the assessment of library similarity was analogous to that of DiverseSolutions in terms of cell occupancy, a key difference in Geysen et al.'s method is that there were multiple points for each scaffold corresponding to multiple scaffold conformations rather than the multiple points that result from multiple compounds in diversity space. Orientation/rotation in space was also considered for each scaffold since this may affect the presentation of the triangle with respect to a receptor. Examination of the regions of the cell based diversity space produced by a library comparison, makes it possible to find regions for "surrogate synthesis" or, in other words, to identify the optimal library to synthesize from a structurally related set of libraries. For a set of libraries that access the same diversity space, it is desirable to select the library that maps the most information about binding with regard to all the members of the set. This selection of the best "surrogate" library is effectively a form of experimental design – the authors were endeavoring to find a

representative subset that best maps the whole. Finally, since the method is independent of library size, it becomes possible to use this method to scaffold hop based on a single compound. Most recently, Geysen et al. demonstrated their approach through an analysis of Dolle's annual collections of combinatorial libraries [200–210] which yielded proposals for scaffold hopping and optimal library design opportunities [213].

The commercially available scaffold hopping method, SHOP [214] available from Molecular Design Ltd [215] or from Molecular Discovery Ltd [216] also makes use of Dolle's compilations of libraries, but for the specific purpose of generating synthetically accessible fragment libraries for use in bio-isostere and scaffold searching. SHOP, employs a GRID [217] based method to search scaffold databases using three types of 3D-descriptors [214]. The procedure compares the similarity of the 3D structure of a query scaffold to those in the database in order to find substitutes that retain the geometry, shape and interaction patterns of the query. Conformation generation is performed on the query and the database contains pre-calculated conformers for each scaffold. The descriptors used are specific to the attachment or anchor points where R-groups would be attached. Distances and dihedral angles are calculated between these anchor points. Distances are recorded in 0.4 Å bins and the dihedrals are transformed into binned distances of 5°. Both are described as Gaussian functions to overlap between different bins in the similarity analysis. In addition, the shape and interaction pattern with respect to the sites is encoded. Shape is based on frequency analysis between each anchor point and the scaffold surface computed with molecular interaction fields (MIFS) using the GRID program [217] (which is called automatically within SHOP) to calculate interaction energies in a 5 Å box-shaped grid around the scaffold. The distances between each anchor point and each grid point with an energy value greater than one kcal/mol determined using a non-polar charged probe are counted. The shape description is created by binning the number of distances as a function of the distance in 0.4 Å wide bins. The third type of descriptors is based on GRIND or "GRID alignment independent descriptors" [218]. Unlike Geysen's methodology above, this method assumes that the scaffold will have interactions with the receptor rather than just presenting groups to interact with it. These potential interactions are estimated as favorable interaction energies calculated between the scaffold and a GRID-probe and are recorded as a function of the distances between the fixed anchor points and the interaction points. Five probes from GRID are used: DRY (hydrophobic), N1(hydrogen bond donor), O (hydrogen bond acceptor), N1+ (positive charge) and O− (negative charge). SHOP uses a grid step of 0.5 Å while other GRID parameters are kept as defaults. The distances between the anchor points and the MIFs are

computed and the corresponding energy values are binned according to the distances. Bergmann et al. [214] validated the SHOP method through a statistical experimental design approach using a database spiked with fragments of known ligands for three different protein targets. SHOP found 5/8 thrombin scaffolds within the top ten ranked scaffolds and 7/8 within the top 31. All seven HIV protease scaffolds were found within the top 10 and all 31 neuraminidase scaffolds were found in the top 31 ranked scaffolds. SHOP was also able to identify new scaffolds possessing significantly different chemotypes from the queries that were predicted by docking with GLIDE [87, 219] to have similar binding modes to the query scaffolds.

The Bergmann study was carried out using the ligands from various X-ray crystal structures to derive the query scaffolds, but SHOP2.0 and subsequent versions also can be used to search for scaffolds or elaborate the R-positions of a scaffold using a crystal structure. The program generates pharmacophore "site points" derived from the residues in the active site to create MIP based fingerprints for searching the database. R-groups found by the search can also be automatically combined with the user provided scaffold query for an enumerated library. Also included with the program are scripts to generate fragment databases using commercial (or other) reagent databases plus reaction files. An automatic fragment naming scheme allows the user to trace fragment or scaffold hits back to both the original reagent and the reaction that could be used to place the fragment on the scaffold [220]. The number of attachment points, unlike the Geysen method, can vary from one to six and makes the method suitable for use in the design of both combinatorial and parallel libraries.

Related to SHOP is the program FLAP [122, 221] which was mentioned above in the context of pharmacophores. This program also uses GRID probes to generate site point based fingerprints for virtual screening either using an active site based query or a ligand based query. In this instance, rather than using databases, the user generates conformations and their pharmacophore site point fingerprints on the fly for the ligands of interest. Since the fingerprint based method is extremely fast – much faster than docking – it is especially appropriate for the screening of enumerated virtual libraries for design purposes.

Another approach which has been proposed for scaffold hopping using a receptor has already been discussed in the context of structure based library design. Shelley presented CombiGLIDE in this context [162] but no prospective studies have yet appeared.

An alternative approach to SHOP and FLAP is found in the OpenEye Scientific Software program EON [222]. Jennings and Tennant [223] used this program and ROCS [212] to examine the use of shape and electrostatic similarity. They calculated an electrostatic Tanimoto by superimposing molecules with ROCS,

determined their electrostatic potentials with the ZAP Poisson–Boltzmann solver and scored the overlays with EON. In their examination of two small sets of molecules including a linker set and a set of compounds from the Maybridge database [224] of commercially available compounds, they compared the EON based similarities with 2D fingerprints and demonstrated that shape and electrostatic field similarity metrics capture relevant molecular similarities. The 2D fingerprints, though often used for library design, were not successful in doing so.

An adaptation of DOCK [158] for scaffold hopping and bioisostere searching, KIN, was published by Andrew Good [225] and performed shape based searching based on STO-3G Gaussian function derived electron density. The code also allowed the use of exclusion regions based on Gaussian shape overlap and the use of pharmacophore constraints. The clique searching feature within DOCK was employed to allow each atom in the query template to be assigned explicitly to a particular chemical type. Additionally, it was possible to define query atoms as critical points or define critical regions. This allowed "exquisite" control over both what chemical features existed in the query and over which atoms or groups of atoms were mapped. For KIN validation, Good employed the dataset previously used to validate ROCS [212]. While not explicitly discussed in the publication, the method should also be applicable to the use of queries defined from a protein receptor site. The scoring function for KIN is highly flexible and author is currently exploring enhancements such as R-group mapping for de novo design, incorporation of pharmacophore constraints into the Gaussian functions weighting, chemical (color) scoring and individual weightings for exclusion spheres, the addition of other Gaussian properties such as MEPs (Maps of electrostatic potential) and template specific atom types to decouple color scoring from pharmacophore constraints.

2.5. Other Library Design Trends

Synthetic accessibility and truly practical reagent selection have become key areas for future library design methods development. In addition to Zamora's SHOP databases mentioned above, Cramer [226] has published Tripos' AllChem project which is intended to generate vast numbers of synthetically accessible structures. The project employs multiple property optimizations including cost/benefit analysis along with synthetic routes to designed structures. This is described by the authors as a work in progress and includes scaffold hopping methodology. The system is intended to generate a vast number of synthetically accessible and medicinally relevant structures for searching.

The emphasis of Allchem on generating huge numbers of structures raises important issues for scaffold hopping. Many of the scaffold hopping tools require databases of multiple conformations for effective scaffold hopping. Efficient database design

can be a problem – is it necessary to include every possible ring system with every possible decoration at positions not specified as attachment points or is a more selective subset based on ring systems and perhaps a limited number of more flexible structures sufficient? The former could certainly be generated through complete deconstruction of a corporate or commercial database but the size would probably preclude extensive conformational searching or storage unless a cpu grid were employed. Both speed and data storage requirements are concerns. Analysis of the ranked results from scaffold hopping searches across conformations requires significant filtering to remove redundant answers, even with smaller databases. Scoring functions for scaffolds and bioisosteres will undoubtedly present similar challenges to those found by docking software developers. These are all issues which remain to be addressed.

A similar filtering issue applies to reagents or side chains for combinatorial and parallel library designs. Is it necessary to mine the thousands of commercially available amines to design a library or is there a reasonable subset that can be mined according to the reactions for which they will be employed? Truchon and Bayley [227, 228] clearly believed that filtering and pruning of reagent sets for optimal library design was appropriate and necessary. To this end they developed an algorithm which employed a deterministic procedure based on product properties, but without explicit enumeration, that took reagent lists from the ACD and produced "cleaned" reagent lists that were both more tractable in total reagent number and enriched in reagents that led to products which satisfied pre-established "goodness" criteria.

3. Conclusions

The trend in library design in the twenty-first century has been toward smaller, focused, information driven libraries in both combinatorial and array formats. Full combinatorial matrices have been displaced by more formats such as sparse matrices that exclude compounds with undesirable properties. While the libraries and arrays that are actually synthesized have grown smaller, computing grids and cpu clusters have facilitated the development of ligand and receptor structure based tools that screen virtual libraries of hundreds of thousands of molecules during the design process. In the early days, large diverse libraries were intended to be screened via HTS in a paradigm that was more than slightly analogous to the agricultural industry's 1980s "spray and pray" style of herbicide discovery. Today, library design employs a collection of tools and methods that incorporate available information on receptors, structure activity relationships, ADMET, etc.

Library size is scaled to fit the intended use of the library, be that corporate deck enhancement, lead discovery, SAR determination or single target optimization. Library size is no longer merely a function of the available automated synthesis tools. Currently, design methods incorporate available information about ligand potency and where possible, receptor structure. Tools have recently been developed to incorporate bio-isosteres, scaffold-hopping and synthetic feasibility. As a result, the field of library design has now come of age and promises to truly impact drug discovery.

References

1. Burbaum, J. J., Ohlmeyer, M. H. J., Reader, J. C., Henderson, I., Dillard, L. W., Li, G., Randle, T. L., Sigal, N. H., Chelsky, D., and Baldwin, J. J. (1995) A paradigm for drug discovery employing encoded combinatorial libraries *Proc. Natl. Acad. Sci. U.S.A.* **92**(13), 6027–6031
2. Blondelle, S. E., Crooks, E., Ostresh, J. M., and Houghten, R. A. (1999) Mixture-based heterocyclic combinatorial positional scanning libraries: Discovery of bicyclic guanidines having potent antifungal activities against *Candida albicans* and *Cryptococcus neoformans Antimicrob. Agents Chemother.* **43**(1), 106–114
3. Boger, D. L., Jiang, W., and Goldberg, J. (1999) Convergent solution-phase combinatorial synthesis with multiplication of diversity through rigid biaryl and diarylacetylene couplings *J. Org. Chem.* **64**(19), 7094–7100
4. Ferry, G., Boutin, J. A., Atassi, G., Fauchere, J. L., and Tucker, G. C. (1997) Selection of a histidine-containing inhibitor of gelatinases through deconvolution of combinatorial tetrapeptide libraries *Mol. Divers.* **2**(3), 135–146
5. Lutzke, R. A. P., Eppens, N. A., Weber, P. A., Houghten, R. A., and Plasterk, R. H. A. (1995) Identification of a hexapeptide inhibitor of the human immunodeficiency virus integrase protein by using a combinatorial chemical library *Proc. Natl. Acad. Sci. U.S.A.* **92**(25), 11456–11460
6. Samson, I., Kerremans, L., Rozenski, J., Samyn, B., Van Beeumen, J., Van Aerschot, A., and Herdewijn, P. (1995) Identification of a peptide inhibitor against glycosomal phosphoglycerate kinase of *Trypanosoma brucei* by a synthetic peptide library approach *Bioorg. Med. Chem.* **3**(3), 257–265
7. Bures, M. G., and Martin, Y. C. (1998) Computational methods in molecular diversity and combinatorial chemistry *Curr. Opin. Chem. Biol.* **2**(3), 376–380
8. Van Drie, J. H., and Lajiness, M. S. (1998) Approaches to virtual library design *Drug Discov. Today* **3**(6), 274–283
9. Spellmeyer, D. C., and Grootenhuis, P. D. J. (1999) Recent developments in molecular diversity. Computational approaches to combinatorial chemistry *Annu. Rep. Med. Chem.* **34**, 287–296
10. Drewry, D. H., and Stanley Young, S. (1999) Approaches to the design of combinatorial libraries *Chemom. Intell. Lab. Syst.* **48**(1), 1–20
11. Leach, A. R., and Hann, M. M. (2000) The in silico world of virtual libraries *Drug Discov. Today* **5**(8), 326–336
12. Lewis, R. A., Pickett, S. D., and Clark, D. E. (2000) Computer-aided molecular diversity analysis and combinatorial library design *Rev. Comput. Chem.* **16**, 1–51
13. Agrafiotis, D. K., Myslik, J. C., and Salemme, F. R. (1999) Advances in diversity profiling and combinatorial series design *Annu. Rep. Comb. Chem. Mol. Divers.* **2**, 71–92
14. Willett, P. (2000) Chemoinformatics – Similarity and diversity in chemical libraries *Curr. Opin. Biotechnol.* **11**(1), 85–88
15. Engels, M. F. M., and Venkatarangan, P. (2001) Smart screening: Approaches to efficient HTS *Curr. Opin. Drug Discov. Devel.* **4** (3), 275–283
16. Andersson, P. M., Sjostrom, M., Wold, S., and Lundstedt, T. (2001) Strategies for subset selection of parts of an in-house chemical library *J. Chemom.* **15**(4), 353–369
17. Andersson, P. M., Linusson, A., Wold, S., Sjostrom, M., Lundstedt, T., and Norden, B. (1999) Design of small libraries for lead exploration *Mol. Divers. Drug Des.* 197–220
18. Linusson, A., Gottfries, J., Lindgren, F., and Wold, S. (2000) Statistical molecular design

of building blocks for combinatorial chemistry *J. Med. Chem.* **43**(7), 1320–1328
19. Brannigan, L. H., Grieshaber, M. V., and Schnur, D. M. (1995) Experimental design in organic synthesis *ACS Symp. Ser.* **589** (Computer-Aided Molecular Design), 225–235
20. Schnur, D. (1999) Design and diversity analysis of large combinatorial libraries using cell-based methods *J. Chem. Inf. Comput. Sci.* **39**(1), 36–45
21. Schnur, D. M., and Venkatrangan, P. (2001) Applications of cell-based diversity to combinatorial library design. In: Ghose, A. K. (ed). *Combinatorial Design and Evaluation*, Marcel Dekker, New York, NY
22. Olsson, T., and Oprea, T. I. (2001) Cheminformatics: A tool for decision-makers in drug discovery *Curr. Opin. Drug Discov. Devel.* **4** (3), 308–313
23. Xue, L., Stahura, F. L., and Bajorath, J. (2004) Cell-based partitioning *Methods Mol. Biol.* **275**, 279–289
24. Pearlman Robert, S. DiverseSolutions Manual, Laboratory for Molecular Graphics and Theoretical Modeling, College of Pharmacy, University of Texas at Austin, Austin TX 78712
25. Cavallaro, C. L., Schnur, D.M., Johnson, S., and Tebben, A. Manuscript in preparation
26. van de Waterbeemd, H., Smith, D. A., Beaumont, K., and Walker, D. K. (2001) Property-based design: Optimization of drug absorption and pharmacokinetics *J. Med. Chem.* **44**(9), 1313–1333
27. Palm, K., Stenberg, P., Luthman, K., and Artursson, P. (1997) Polar molecular surface properties predict the intestinal absorption of drugs in humans *Pharm. Res.* **14**(5), 568–571
28. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings *Adv. Drug Deliv. Rev.* **23**(1–3), 3–25
29. Oprea, T. I. (2002) Current trends in lead discovery: Are we looking for the appropriate properties? *J. Comput. Aided Mol. Des.* **16**(5/6), 325–334
30. Oprea, T. I., Zamora, I., and Ungell, A. -L. (2002) Pharmacokinetically based mapping device for chemical space navigation *J. Comb. Chem.* **4**(4), 258–266
31. Oprea, T. I., and Gottfries, J. (2000) Toward minimalistic modeling of oral drug absorption1 *J. Mol. Graph. Model.* **17**(5/6), 261–274
32. Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., Lundell, G. F., Veber, D. F., Anderson, P. S., et al. (1988) Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **31**(12), 2235–2246
33. Patchett, A. A., and Nargund, R. P. (2000) Chapter 26. Privileged structures – An update *Annu. Rep. Med. Chem.* **35**, 289–298
34. Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., and Labaudiniere, R. F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures *J. Med. Chem.* **42**(17), 3251–3264
35. Lamb, M. L., Bradley, E. K., Beaton, G., Bondy, S. S., Castellino, A. J., Gibbons, P. A., Suto, M. J., and Grootenhuis, P. D. J. (2004) Design of a gene family screening library targeting G-protein coupled receptors *J. Mol. Graph. Model.* **23**(1), 15–21
36. Savchuk, N. P., Tkachenko, S. E., and Balakin, K. V. (2005) Rational design of GPCR-specific combinatorial libraries based on the concept of privileged substructures *Methods Princ. Med. Chem.* **23**(Chemoinformatics in Drug Discovery), 287–313
37. Lowrie, J. F., Delisle, R. K., Hobbs, D. W., and Diller, D. J. (2004) The different strategies for designing GPCR and kinase targeted libraries *Comb. Chem. High Throughput Screen.* **7**(5), 495–510
38. Prien, O. (2005) Target-family-oriented focused libraries for kinases-conceptual design aspects and commercial availability *Chembiochem* **6**(3), 500–505
39. Stewart, E. L., Brown, P. J., Bentley, J. A., and Willson, T. M. (2001) Selection, application, and validation of a set of molecular descriptors for nuclear receptor ligands *Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26–30, 2001*, COMP-182
40. Lang, S. A., Kozyukov, A. V., Balakin, K. V., Skorenko, A. V., Ivashchenko, A. A., and Savchuk, N. P. (2003) Classification scheme for the design of serine protease targeted compound libraries *J. Comput. Aided Mol. Des.* **16**(11), 803–807
41. Schuffenhauer, A., Zimmermann, J., Stoop, R., Van der Vyver, J. -J., Lecchini, S., and Jacoby, E. (2002) An ontology for pharmaceutical ligands and its application

for in silico screening and library design *J. Chem. Inf. Comput. Sci.* **42**(4), 947–955
42. Aureus Pharmaceuticals, 174, Quai de Jemmapes, 75010 Paris, France
43. Jubilant Biosys, Ltd.,8575 Window Latch Way, Columbia, MD 21045
44. Sertanty Inc., 1735 N. First St. #102, San Jose CA, 95112
45. McGregor, M. J., and Muskal, S. M. (2000) Pharmacophore fingerprinting. 2. Application to primary library design *J. Chem. Inf. Comput. Sci.* **40**(1), 117–125
46. Naumann, T., and Matter, H. (2002) Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: Target family landscapes *J. Med. Chem.* **45**(12), 2366–2378
47. Rad, R., Mracec, M., Mracec, M., and Oprea, T. (2007) The privileged structures hypothesis for G protein-coupled receptors – Some preliminary results *Rev. Roum. Chim.* **52**(8–9), 853–858
48. Klabunde, T., and Hessler, G. (2002) Drug design strategies for targeting G-protein-coupled receptors *Chembiochem* **3**(10), 928–944
49. Gooding, O. W. (2004) Process optimization using combinatorial design principles: Parallel synthesis and design of experiment methods *Curr. Opin. Chem. Biol.* **8**(3), 297–304
50. Debnath, A. K. (2001) Quantitative structure–activity relationship (QSAR), a versatile tool in drug design. In: Ghose, A. K. (ed). *Combinatorial Library Design and Evaluation*, Marcel Dekker, Inc, New York, New York
51. Oprea, T. I., Davis, A. M., Teague, S. J., and Leeson, P. D. (2001) Is there a difference between leads and drugs? A historical perspective *J. Chem. Inf. Comput. Sci.* **41**(5), 1308–1315
52. Teague, S. J., Davis, A. M., Leeson, P. D., and Oprea, T. (1999) The design of leadlike combinatorial libraries *Angew. Chem. Int. Ed.* **38**(24), 3743–3748
53. Hansch, C., and Albert, L. (1995) Calculation of octanol-water partition coefficients by fragments. In: Heller, S. (ed). *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, DC
54. Hansch, C., Hoekman, D., and Gao, H. (1996) Comparative QSAR: Toward a deeper understanding of chemicobiological interactions *Chem. Rev.* **96**(3), 1045–1075
55. Oprea, T. I., Olah, M., Mracec, M., Rad, R., Ostopovici, L., Bora, A., Hadaruga, N., and Bologa, C. G. (2005) Mapping bioactivity space for fragment-based lead discovery *Abstracts of Papers, 229th ACS National Meeting, San Diego, CA, United States, March 13–17, 2005*, MEDI-277
56. Oprea, T. I. (2007) Lead-like, drug-like or "pub-like": How different are they? *Abstracts of Papers, 233rd ACS National Meeting, Chicago, IL, United States, March 25–29, 2007*, CINF-014
57. Oprea, T. I. (2005) Pursuing leadlikeness in pharmaceutical research *Jt. Meet. Med. Chem., Proc., Vienna, Austria, June 20–23, 2005*, 1–4
58. Jaakola, V. -P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y. T., Lane, J. R., Ijzerman, A. P., and Stevens, R. C. (2008) The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist *Science* **322**(5905), 1211–1217
59. Scheerer, P., Park, J. H., Hildebrand, P. W., Kim, Y. J., Krauss, N., Choe, H. -W., Hofmann, K. P., and Ernst, O. P. (2008) Crystal structure of opsin in its G-protein-interacting conformation *Nature* **455** (7212), 497–502
60. Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G. W., Tate, C. G., and Schertler, G. F. X. (2008) Structure of a b1-adrenergic G-protein-coupled receptor *Nature* **454**(7203), 486–491
61. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. -W., and Ernst, O. P. (2008) Crystal structure of the ligand-free G-protein-coupled receptor opsin *Nature* **454**(7201), 183–187
62. Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., Choi, H. -J., Kuhn, P., Weis, W. I., Kobilka, B. K., Stevens, R. C., Takeda, S., Kadowaki, S., Haga, T., Takaesu, H., Mitaku, S., Fredriksson, R., Lagerstrom, M. C., Lundin, L. G., Schioth, H. B., Pierce, K. L., Premont, R. T., Lefkowitz, R. J., Lefkowitz, R. J., Shenoy, S. K., and Rosenbaum, D. M. (2007) High-resolution crystal structure of an engineered human b2-adrenergic G protein-coupled receptor *Science* **318** (5854), 1258–1265
63. Rasmussen, S. G. F., Choi, H. -J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R. P., Sanishvili, R., Fischetti, R. F., Schertler, G. F. X., Weis, W. I., and Kobilka, B. K. (2007) Crystal structure of the human b2 adrenergic G-protein-coupled receptor *Nature* **450**(7168), 383–387
64. Worth, C. L., Kleinau, G., and Krause, G. (2009) Comparative sequence and structural analyses of G-protein-coupled receptor

crystal structures and implications for molecular models *PLoS One* **4**(9)
65. Mobarec, J. C., Sanchez, R., and Filizola, M. (2009) Modern homology modeling of g-protein coupled receptors: Which structural template to use? *J. Med. Chem.* **52** (16), 5207–5216
66. Panigrahi, S. K., and Desiraju, G. R. (2003) Homology modelling in protein structure prediction: Epidermal growth factor receptor kinase domain *Natl. Acad. Sci. Lett.* **27** (1 & 2), 1–11
67. Mahadevan, D., Bearss David, J., and Vankayalapati, H. (2003) Structure-based design of novel anti-cancer agents targeting aurora kinases *Curr. Med. Chem. Anticancer Agents* **3**(1), 25–34
68. Mozzicafreddo, M., Cuccioloni, M., Cecarini, V., Eleuteri, A. M., and Angeletti, M. (2009) Homology modeling and docking analysis of the interaction between polyphenols and mammalian 20S proteasomes *J. Chem. Inf. Model.* **49**(2), 401–409
69. Zhou, H., Singh, N. J., and Kim, K. S. (2006) Homology modeling and molecular dynamics study of West Nile virus NS3 protease: A molecular basis for the catalytic activity increased by the NS2B cofactor *Proteins* **65**(3), 692–701
70. Majer, F., Pavlickova, L., Majer, P., Hradilek, M., Dolejsi, E., Hruskova-Heidingsfeldova, O., and Pichova, I. (2006) Structure-based specificity mapping of secreted aspartic proteases of *Candida parapsilosis*, *Candida albicans*, and *Candida tropicalis* using peptidomimetic inhibitors and homology modeling *Biol. Chem.* **387**(9), 1247–1254
71. ClogP is available from Daylight Information Systems, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA, 92691
72. Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985) Atom pairs as molecular features in structure–activity studies: Definition and applications *J. Chem. Inf. Comput. Sci.* **25**(2), 64–73
73. Daylight Information Systems, Inc., 27401 Los Altos, Suite 360, Mission Viejo, CA, 92691
74. McGregor, M. J., and Muskal, S. M. (1999) Pharmacophore fingerprinting. 1. Application to QSAR and focused library design *J. Chem. Inf. Comput. Sci.* **39**(3), 569–574
75. Beno, B. R., and Mason, J. S. (2001) The design of combinatorial libraries using properties and 3D pharmacophore fingerprints *Drug Discov. Today* **6**(5), 251–258
76. Pearlman, R. S., and Smith, K. M. (1998) Software for chemical diversity in the context of accelerated drug discovery *Drugs Future* **23**(8), 885–895
77. Pearlman, R. S., and Smith, K. M. (1999) Metric validation and the receptor-relevant subspace concept *J. Chem. Inf. Comput. Sci.* **39**(1), 28–35
78. Kubinyi, H. (2004) 2D QSAR models: Hansch and Free-Wilson analyses *Comput. Med. Chem. Drug Discovery*, 539–570
79. Topliss, J. G. (1993) Some observations on classical QSAR *Perspect. Drug Discovery Des.* **1**(2), 253–268
80. Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., and Moos, W. H. (1995) Measuring diversity: Experimental design of combinatorial libraries for drug discovery *J. Med. Chem.* **38**(9), 1431–1436
81. Austel, V. (1995) Experimental design in synthesis planning and structure–property correlations. Experimental design *Methods Princ. Med. Chem.* **2**, 49–62
82. Brannigan, L. H., Grieshaber, M. V., and Schnur, D. M. (1995) Use experimental design in organic synthesis *Chem. Tech.* **25**(3), 29–35
83. Rose, S., and Stevens, A. (2003) Computational design strategies for combinatorial libraries *Curr. Opin. Chem. Biol.* **7**(3), 331–339
84. Mitchell, T., and Showell, G. A. (2001) Design strategies for building drug-like chemical libraries *Curr. Opin. Drug Discov. Devel.* **4**(3), 314–318
85. Gillet, V. J. (2002) Reactant- and product-based approaches to the design of combinatorial libraries *J. Comput. Aided Mol. Des.* **16** (5/6), 371–380
86. Pearlman Robert, S. DiverseSolutions. Is available from Tripos, 1699 South Hanley Road, St. Louis, MO, 63144, http://www.tripos.com
87. GLIDE and CombiGLIDE are modules of the MAESTRO software package available from Schrödinger LLC, 120 West 45th Street, 17th Floor, New York, NY 10036. http://www.schrodinger.com
88. Pearlman Robert, S. Benchware (LibraryMaker/LibraryDesigner) was available from Tripos, 1699 South Hanley Road, St. Louis, MO, 63144, http://www.tripos.com
89. Matter, H. (1997) Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors *J. Med. Chem.* **40**(8), 1219–1229
90. Zuccotto, F. (2003) Pharmacophore features distributions in different classes of

compounds *J. Chem. Inf. Comput. Sci.* **43**(5), 1542–1552
91. Hall, L. H., and Kier, L. B. (1990) Determination of topological equivalence in molecular graphs from the topological state *Quant. Struct.-Act. Relat.* **9**(2), 115–131
92. Hall, L. H., Mohney, B., and Kier, L. B. (1991) The electrotopological state: An atom index for QSAR *Quant. Struct.-Act. Relat.* **10**(1), 43–51
93. Hassan, M., Bielawski, J. P., Hempel, J. C., and Waldman, M. (1996) Optimization and visualization of molecular diversity of combinatorial libraries *Mol. Divers.* **2**(1/2), 64–74
94. Cerius2 was available from Accelrys, Inc., 9685 Scranton Road, San Diego, CA 92121-3752, but has since been incorporated into Discovery Studio. http://accelrys.com
95. Jamois, E. A., Lin, C. T., and Waldman, M. (2003) Design of focused and restrained subsets from extremely large virtual libraries *J. Mol. Graph. Model.* **22**(2), 141–149
96. Brown, R. D., Hassan, M., and Waldman, M. (2004) Tools for designing diverse, druglike, cost-effective combinatorial libraries. In: Bajorath, J. (ed). *Chemoinformatic: Concepts, Methods and Tools for Drug Discovery*, first Ed., Humana Press, Totowa, NJ
97. Linusson, A., Gottfries, J., Olsson, T., Oernskov, E., Folestad, S., Norden, B., and Wold, S. (2001) Statistical molecular design, parallel synthesis, and biological evaluation of a library of thrombin inhibitors *J. Med. Chem.* **44**(21), 3424–3439
98. MODDE is available form Umetrics, Inc., Kinnelon, N. J., Kinnelon, N. J. http://www.umetrics.com/
99. Lee, A., and Breitenbucher, J. G. (2003) The impact of combinatorial chemistry on drug discovery *Curr. Opin. Drug Discov. Devel.* **6**(4), 494–508
100. Skiles, J. W., Gonnella, N. C., and Jeng, A. Y. (2001) The design, structure, and therapeutic application of matrix metalloproteinase inhibitors *Curr. Med. Chem.* **8**(4), 425–474
101. Pauls, H. W., and Ewing, W. R. (2001) The design of competitive, small-molecule inhibitors of coagulation factor Xa. *Curr. Top. Med. Chem.* **1**(2), 83–100
102. Pearlman, R. S., and Smith, K. M. (1998) Novel metrics and validation of metrics for chemical diversity *Alfred Benzon Symp.* **42** (Rational Molecular Design in Drug Research), 165–185
103. Pearlman, R. S., and Smith, K. M. (1999) Novel algorithms for the design of diverse and focussed combinatorial libraries *Book of Abstracts, 217th ACS National Meeting, Anaheim, Calif., March 21–25,* COMP-197
104. Wang, X. -C., and Saunders, J. (2001) GPCR library design *Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26–30, 200*, MEDI-012
105. Mason, J. S., and Beno, B. R. (2000) Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: Simultaneous optimization and structure-based diversity *J. Mol. Graph. Model.* **18**(4/5), 438–451
106. Pirard, B., and Pickett, S. D. (2000) Classification of kinase inhibitors using BCUT descriptors *J. Chem. Inf. Comput. Sci.* **40**(6), 1431–1440
107. Manallack, D. T., Pitt, W. R., Gancia, E., Montana, J. G., Livingstone, D. J., Ford, M. G., and Whitley, D. C. (2002) Selecting screening candidates for kinase and G protein-coupled receptor targets using neural networks *J. Chem. Inf. Comput. Sci.* **42**(5), 1256–1262
108. Mason, J. S., Good, A. C., and Martin, E. J. (2001) 3-D pharmacophores in drug discovery *Curr. Pharm. Des.* 7(7), 567–597
109. Good, A. C., Cho, S. -J., and Mason, J. S. (2004) Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening *J. Comput. Aided Mol. Des.* **18**(7–9), 523–527
110. Bradley, E. K., Beroza, P., Penzotti, J. E., Grootenhuis, P. D. J., Spellmeyer, D. C., and Miller, J. L. (2000) A rapid computational method for lead evolution: Description and application to a1-adrenergic antagonists *J. Med. Chem.* **43**(14), 2770–2774
111. Lanctot, J. K., Putta, S., Lemmen, C., and Greene, J. (2003) Using ensembles to classify compounds for drug discovery *J. Chem. Inf. Comput. Sci.* **43**(6), 2163–2169
112. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching *J. Chem. Inf. Comput. Sci.* **38**(6), 983–996
113. Lanctot, J. K., Srinivasan, J., Lamb, M. L., Lemmen, C., Eksterowicz, J. E., and Putta, S. (2003) Mitigating the sensitivity of pharmacophoric fingerprints *Abstracts of Papers, 225th ACS National Meeting, New Orleans, LA, United States, March 23–27, 2003*, COMP-323
114. Pickett, S. D., McLay, I. M., and Clark, D. E. (2000) Enhancing the hit-to-lead

properties of lead optimization libraries *J. Chem. Inf. Comput. Sci.* **40**(2), 263–272
115. Zuckermann, R. N., Martin, E. J., Spellmeyer, D. C., Stauber, G. B., Shoemaker, K. R., Kerr, J. M., Figliozzi, G. M., Goff, D. A., Siani, M. A., Simon, R., Banville, S. C., Brown, E. G., Wang, L., Richter, L. S., and Moos, W. H. (1994) Discovery of nanomolar ligands for 7-transmembrane G-protein-coupled receptors from a diverse N-(substituted)glycine peptoid library *J. Med. Chem.* **37**(17), 2678–2685
116. Bradley, E. K., Miller, J. L., Saiah, E., and Grootenhuis, P. D. J. (2003) Informative library design as an efficient strategy to identify and optimize leads: Application to cyclin-dependent kinase 2 antagonists *J. Med. Chem.* **46**(20), 4360–4364
117. Deanda, F., Stewart, E. L., Reno, M. J., and Drewry, D. H. (2008) Kinase-targeted library design through the application of the Pharmprint methodology *J. Chem. Inf. Model.* **48**(12), 2395–2403
118. Deanda, F., and Stewart, E. L. (2004) Application of the PharmPrint methodology to two protein kinases *J. Chem. Inf. Comput. Sci.* **44**(5), 1803–1809
119. Mason, J. S., and Cheney, D. L. (1999) Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores *Pac. Symp. Biocomput. '99, Mauna Lani, Hawaii, Jan. 4–9, 1999*, 456–467
120. Murray, C. M., and Cato, S. J. (1999) Design of libraries to explore receptor sites *J. Chem. Inf. Comput. Sci.* **39**(1), 46–50
121. Arnold, J. R., Burdick, K. W., Pegg, S. C. H., Toba, S., Lamb, M. L., and Kuntz, I. D. (2004) SitePrint: Three-dimensional pharmacophore descriptors derived from protein binding sites for family based active site analysis, classification, and drug design *J. Chem. Inf. Comput. Sci.* **44**(6), 2190–2198
122. Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F., and Mason, J. S. (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): Theory and application *J. Chem. Inf. Model.* **47**(2), 279–294
123. Lamb, M. L., Bradley, E. K., Spellmeyer, D. C., Suto, M. J., and Grootenhuis, P. D. J. (2001) Iterative design of gene-family directed screening libraries *Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26–30, 2001*, COMP-053
124. Grootenhuis, P. D. J., Lamb, M. L., Bradley, E. K., Myers, P. L., Shirley, W. A., Rogers, D., Castellino, A. J., and Miller, J. L. (2003) Process for the informative and iterative design of a gene-family screening library. In: (Deltagen Research Laboratories, L.L. C., USA). Application: WO
125. Ugi, I., and Steinbrucker, C. (1961) Isonitriles. II. Reaction of isonitriles with carbonyl compounds, amines, and hydrazoic acid *Chem. Ber.* **94**, 734–742
126. Formerly MDL Drug Data Report, MDDR is produced by Symyx and Prous Science. http://www.symyx.com/products/databases/bioactivity/mddr
127. Good, A. C., and Lewis, R. A. (1997) New methodology for profiling combinatorial libraries and screening sets: Cleaning up the design process with HARPick *J. Med. Chem.* **40**(24), 3926–3936
128. Penzotti, J. E., Lamb, M. L., Evensen, E., and Grootenhuis, P. D. J. (2002) A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein *J. Med. Chem.* **45**(9), 1737–1740
129. Ekins, S., Bravi, G., Binkley, S., Gillespie, J. S., Ring, B. J., Wikel, J. H., and Wrighton, S. A. (1999) Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors *J. Pharmacol. Exp. Ther.* **290**(1), 429–438
130. Ekins, S., Bravi, G., Wikel, J. H., and Wrighton, S. A. (1999) Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates *J. Pharmacol. Exp. Ther.* **291**(1), 424–433
131. Sciabola, S., Morao, I., and de Groot, M. J. (2007) Pharmacophoric fingerprint method (TOPP) for 3D-QSAR modeling: Application to CYP2D6 metabolic stability *J. Chem. Inf. Model.* **47**(1), 76–84
132. Johnson, S. R., Yue, H., Conder, M. L., Shi, H., Doweyko, A. M., Lloyd, J., and Levesque, P. (2007) Estimation of hERG inhibition of drug candidates using multivariate property and pharmacophore SAR *Bioorg. Med. Chem.* **15**(18), 6182–6192
133. Brady, G. P., and Yang, Z. P. (2008) Pharmacophore fingerprints and application to target class modeling *Abstracts of Papers, 235th ACS National Meeting, New Orleans, LA, United States, April 6–10, 2008*, COMP-093
134. Yang, Z., and Brady, G. P. (2008) Applications of target class pharmacophore fingerprint modeling and multi-objective genetic algorithm optimization to large-scale combinatorial library design for corporate compound collection enhancement *Abstracts of Papers, 235th ACS National Meeting, New Orleans, LA, United States, April 6–10, 2008, 2008*, COMP-094

135. Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks *J. Med. Chem.* **39**(15), 2887–2893
136. Mason, J. S., Cheney, D. L., Menard, P. R., and Morize, I. (1997) New pharmacophore-based methods to search, profile and design diverse and biased compound databases and libraries *Book of Abstracts, 214th ACS National Meeting, Las Vegas, NV, September 7–11*, COMP-034
137. Classpharmer is available from Simulations Plus, 42505 10th Street West Lancaster, CA 93534-7059, http://www.simulations-plus.com/
138. Schnur, D. M., Hermsmeier, M. A., and Tebben, A. J. (2006) Are target-family-privileged substructures truly privileged? *J. Med. Chem.* **49**(6), 2000–2009
139. Schnur, D., and Hermsmeier, M. A. (2004) Classpharmer and the quest for privileged substructures *Abstracts of Papers, 228th ACS National Meeting, Philadelphia, PA, United States, August 22–26, 2004*, CINF-080
140. PipelinePilot is available from Accelrys, Inc., 9685 Scranton Road, San Diego, CA 92121-3752., http://accelrys.com/
141. ISIS is available from Symyx, http://www.symyx.com/products/software. ISIS was formly available from MDL. ISIS keys and MACCS keys are equivalent
142. Gianti, E., and Sartori, L. (2008) Identification and selection of "privileged fragments" suitable for primary screening *J. Chem. Inf. Model.* **48**(11), 2129–2139
143. Horton, D. A., Bourne, G. T., and Smythe, M. L. (2003) The combinatorial synthesis of bicyclic privileged structures or privileged substructures *Chem. Rev.* **103**(3), 893–930
144. Bleicher, K. H., Green, L. G., Martin, R. E., and Rogers-Evans, M. (2004) Ligand identification for G-protein-coupled receptors: A lead generation perspective *Curr. Opin. Chem. Biol.* **8**(3), 287–296
145. Merlot, C., Domine, D., Cleva, C., and Church, D. J. (2003) Chemical substructures in drug discovery *Drug Discov. Today* **8**(13), 594–602
146. Formerly MDL Drug Data Report, MDDR is currently produced by Symyx and Prous Science. The 1999 version was available from MDL which was later purchased by Elsevier and subsequently sold to Symyx
147. Tripos, 1699 South Hanley Road, St. Louis, MO, 63144
148. Vaz, R. Private communication
149. Severinsen, R., Bourne, G. T., Tran, T. T., Ankersen, M., Begtrup, M., and Smythe, M. L. (2008) Library of biphenyl privileged substructures using a safety-catch linker approach *J. Comb. Chem.* **10**(4), 557–566
150. Bondensgaard, K., Ankersen, M., Thogersen, H., Hansen, B. S., Wulff, B. S., and Bywater, R. P. (2004) Recognition of privileged structures by g-protein coupled receptors *J. Med. Chem.* **47**(4), 888–899
151. Jimonet, P., and Jaeger, R. (2004) Strategies for designing GPCR-focused libraries and screening sets *Curr. Opin. Drug Discov. Devel.* **7**(3), 325–333
152. Rodrigues de Sa Alves, F., Barreiro, E. J., and Fraga, C. A. M. (2009) From nature to drug discovery: The indole scaffold as a 'privileged structure' *Mini Rev. Med. Chem.* **9**(7), 782–793
153. Lewis, R. A. (2008) Computer-aided drug design 2005–2007 *Chem. Modell.* **5**, 51–66
154. Lang, P. T., Aynechi, T., Moustakas, D., Shoichet, B., Kuntz, I. D., Brooijmans, N., and Oshiro, C. M. (2007) Molecular docking and structure-based design *Drug Discovery Res.* 3–23
155. Abraham, D. J. (2006) Structure-based drug design – A historical perspective and the future *Compr. Med. Chem.* II 4, 65–86
156. Lange, G. (2006) Structure-based drug design – The use of protein structure in drug discovery *Compr. Med. Chem. II* **4**, 597–650
157. Sun, Y., Ewing, T. J. A., Skillman, A. G., and Kuntz, I. D. (1998) CombiDOCK: Structure-based combinatorial docking and library design *J. Comput. Aided Mol. Des.* **12**(6), 597–604
158. Shoichet, B. K., Bodian, D. L., and Kuntz, I. D. (1992) Molecular docking using shape descriptors *J. Comput. Chem.* **13**(3), 380–397
159. Geistlinger, T. R., and Guy, R. K. (2003) Novel selective inhibitors of the interaction of individual nuclear hormone receptors with a mutually shared steroid receptor coactivator 2 *J. Am. Chem. Soc.* **125**(23), 6852–6853
160. Pitts, W. J., Guo, J., Dhar, T. G. M., Shen, Z., Gu, H. H., Watterson, S. H., Bednarz, M. S., Chen, B. -C., Barrish, J. C., Bassolino, D., Cheney, D., Fleener, C. A., Rouleau, K. A., Hollenbaugh, D. L., and Iwanowicz, E. J. (2002) Rapid synthesis of triazine inhibitors of inosine monophosphate dehydrogenase *Bioorg. Med. Chem. Lett.* **12**(16), 2137–2140
161. Sprous, D. G., Lowis, D. R., Leonard, J. M., Heritage, T., Burkett, S. N., Baker, D. S., and Clark, R. D. (2004) OptiDock: Virtual HTS of combinatorial libraries by efficient

sampling of binding modes in product space *J. Comb. Chem.* **6**(4), 530–539
162. Shelley, M., Frye, L. L., Sherman, B. W., Rao, S. N., Beard, H., Mozziconacci, J.-C., and Shenkin, P. S. (2007) New approach to lead optimization and core hopping *Abstracts of Papers, 234th ACS National Meeting, Boston, MA, United States, August 19–23, 2007*, COMP-406
163. Sherman, B. W., Higgs, C., and Shelley, M. (2007) Screening very large virtual libraries using structure-based docking *Abstracts of Papers, 234th ACS National Meeting, Boston, MA, United States, August 19–23, 2007*, COMP-165
164. Shelley, M., Frye, L. L., Murphy, R. B., and Shenkin, P. S. (2006) Focused library design for selective estrogen receptor modulators using CombiGlide *Abstracts of Papers, 232nd ACS National Meeting, San Francisco, CA, United States, Sept. 10–14, 2006*, COMP-040
165. Shenkin, P. S., Frye, L. L., Murphy, R. B., Repasky, M. P., Mainz, D. T., Reboul, M., and Friesner, R. A. (2005) Structure-based design of focused drug-like combinatorial libraries *Abstracts of Papers, 230th ACS National Meeting, Washington, DC, United States, Aug. 28–Sept. 1, 2005*, COMP-117
166. Rarey, M., and Lengauer, T. (2000) A recursive algorithm for efficient combinatorial library docking *Perspect. Drug Discov. Des.* **20**, 63–81
167. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm *J. Mol. Biol.* **261**(3), 470–489
168. Rarey, M., Wefing, S., and Lengauer, T. (1996) Placement of medium-sized molecular fragments into active sites of proteins *J. Comput. Aided Mol. Des.* **10**(1), 41–54
169. Luksch, T., Chan, N. -S., Brass, S., Sotriffer, C. A., Klebe, G., and Diederich, W. E. (2008) Computer-aided design and synthesis of nonpeptidic plasmepsin II and IV inhibitors *ChemMedChem* **3**(9), 1323–1336
170. Gerlach, C., Sohn, C., Craan, T., Diederich, W. E., and Klebe, G. (2006) KNOBLE: KNOwledge-based ligand enumeration *Abstracts of Papers, 232nd ACS National Meeting, San Francisco, CA, United States, Sept. 10–14, 2006*, COMP-246
171. Diederich, W. E., Gerlach, C., Blum, A., Boettcher, J., Brass, S., Luksch, T., and Klebe, G. (2006) Design and synthesis of tailor-made compound libraries via a knowledge-based approach: A case study *Abstracts of Papers, 232nd ACS National Meeting, San Francisco, CA, United States, Sept. 10–14, 2006*, CINF-083
172. Cross, S. S. J. (2005) Improved FlexX docking using FlexS-determined base fragment placement *J. Chem. Inf. Model.* **45**(4), 993–1001
173. Sprous, D., Clark, R., Lowis, D., Leonard, J., and Heritage, T. (2001) Docking combinatorial libraries efficiently using FlexX *Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26–30, 2001*, COMP-033
174. Soltanshahi, F., Liu, Q., and Clark, R. D. (2006) Biasing for favored substituents in kinase library design *Abstracts of Papers, 231st ACS National Meeting, Atlanta, GA, United States, March 26–30, 2006*, COMP-054
175. Schrödinger, https://www.schrodinger.com/Documentation.php?mID=6&sID=24&cID=2150&pdID=2150. In: *CombiGLIDE documentation*, SchrödingerLLC, 120 West 45th Street, 17th Floor, New York, NY 10036
176. Zhou, J. Z. (2008) Structure-directed combinatorial library design *Curr. Opin. Chem. Biol.* **12**(3), 379–385
177. Kolb, P., Kipouros, C. B., Huang, D., and Caflisch, A. (2008) Structure-based tailoring of compound libraries for high-throughput screening: Discovery of novel EphB4 kinase inhibitors. *Proteins* **73**(1), 11–18
178. He, L., Zhang, L., Liu, X., Li, X., Zheng, M., Li, H., Yu, K., Chen, K., Shen, X., Jiang, H., and Liu, H. (2009) Discovering potent inhibitors against the β-hydroxyacyl-acyl carrier protein dehydratase (FabZ) of *Helicobacter pylori*: Structure-based design, synthesis, bioassay, and crystal structure determination *J. Med. Chem.* **52**(8), 2465–2481
179. Chen, G., Zheng, S., Luo, X., Shen, J., Zhu, W., Liu, H., Gui, C., Zhang, J., Zheng, M., Puah, C. M., Chen, K., and Jiang, H. (2005) Focused combinatorial library design based on structural diversity, druglikeness and binding affinity score *J. Comb. Chem.* **7**(3), 398–406
180. Hardcastle, I. R., Ahmed, S. U., Atkins, H., Farnie, G., Golding, B. T., Griffin, R. J., Guyenne, S., Hutton, C., Kaellblad, P., Kemp, S. J., Kitching, M. S., Newell, D. R., Norbedo, S., Northen, J. S., Reid, R. J., Saravanan, K., Willems, H. M. G., and Lunec, J. (2006) Small-molecule inhibitors of the MDM2-p53 protein-protein interaction based on an isoindolinone scaffold *J. Med. Chem.* **49**(21), 6209–6221

181. Stahl, M., Todorov, N. P., James, T., Mauser, H., Boehm, H. -J., and Dean, P. M. (2002) A validation study on the practical use of automated de novo design *J. Comput. Aided Mol. Des.* **16**(7), 459–478
182. Carr, R., and Jhoti, H. (2002) Structure-based screening of low-affinity compounds *Drug Discov. Today* **7**(9), 522–527
183. Hajduk, P. J. (2006) Fragment-based drug design: How big is too big? *J. Med. Chem.* **49**(24), 6972–6976
184. Hajduk, P. J., and Greer, J. (2007) A decade of fragment-based drug design: Strategic advances and lessons learned *Nat. Rev. Drug Discov.* **6**(3), 211–219
185. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings *Adv. Drug Deliv. Rev.* **46**(1–3), 3–26
186. Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003) A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* **8**(19), 876–877
187. Xu, J. (2002) A new approach to finding natural chemical structure classes *J. Med. Chem.* **45**(24), 5311–5320
188. Siegal, G., Ab, E., and Schultz, J. (2007) Integration of fragment screening and library design *Drug Discov. Today* **12**(23 & 24), 1032–1039
189. Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR *Science* **274**(5292), 1531–1534
190. Mayer, M., and Meyer, B. (1999) Characterization of ligand binding by saturation transfer difference NMR spectroscopy *Angew. Chem., Int. Ed.* **38**(12), 1784–1788
191. Dalvit, C., Pevarello, P., Tato, M., Veronesi, M., Vulpetti, A., and Sundstrom, M. (2000) Identification of compounds with binding affinity to proteins via magnetization transfer from bulk water *J. Biomol. NMR* **18**(1), 65–68
192. Vanwetswinkel, S., Heetebrij, R. J., van Duynhoven, J., Hollander, J. G., Filippov, D. V., Hajduk, P. J., and Siegal, G. (2005) TINS, target immobilized NMR screening: An efficient and sensitive method for ligand discovery *Chem. Biol.* **12**(2), 207–216
193. Phillips, K. S., and Cheng, Q. (2007) Recent advances in surface plasmon resonance based techniques for bioanalysis *Anal. Bioanal. Chem.* **387**(5), 1831–1840
194. Rich, R. L., and Myszka, D. G. (2000) Advances in surface plasmon resonance biosensor analysis *Curr. Opin. Biotechnol.* **11**(1), 54–61
195. Neumann, T., Junker, H. D., Schmidt, K., and Sekul, R. (2007) SPR-based fragment screening: Advantages and applications *Curr. Top. Med. Chem.* **7**(16), 1630–1642
196. Graffinity, http://www.graffinity.com
197. Carr, R. A. E., Congreve, M., Murray, C. W., and Rees, D. C. (2005) Fragment-based lead discovery: Leads by design *Drug Discov. Today* **10**(14), 987–992
198. Law, R., Barker, O., Barker, J. J., Hesterkamp, T., Godemann, R., Andersen, O., Fryatt, T., Courtney, S., Hallett, D., and Whittaker, M. (2009) The multiple roles of computational chemistry in fragment-based drug design *J. Comput. Aided Mol. Des.* **23**(8), 459–473
199. Petros, A. M., Dinges, J., Augeri, D. J., Baumeister, S. A., Betebenner, D. A., Bures, M. G., Elmore, S. W., Hajduk, P. J., Joseph, M. K., Landis, S. K., Nettesheim, D. G., Rosenberg, S. H., Shen, W., Thomas, S., Wang, X., Zanze, I., Zhang, H., and Fesik, S. W. (2006) Discovery of a potent inhibitor of the antiapoptotic protein Bcl-xL from NMR and parallel synthesis, *J. Med. Chem.* **49**, 656–663
200. Dolle, R. E. (1997) Comprehensive survey of chemical libraries yielding enzyme inhibitors, receptor agonists and antagonists, and other biologically active agents: 1992 through 1997 *Mol Divers.* **3**(4), 199–233
201. Dolle, R. E. (1998) Comprehensive survey of combinatorial libraries with undisclosed biological activity: 1992–1997. *Mol. Divers.* **4**(4), 233–256
202. Dolle, R. E. (2000) Comprehensive survey of combinatorial library synthesis: 1999 *J. Comb. Chem.* **2**(5), 383–433
203. Dolle, R. E. (2000) Comprehensive survey of combinatorial libraries with undisclosed biological activity: 1992–1997 *Mol. Divers.* **4**(4), 233–256
204. Dolle, R. E. (2001) Comprehensive survey of combinatorial library synthesis: 2000 *J. Comb. Chem.* **3**(6), 477–517
205. Dolle, R. E. (2002) Comprehensive survey of combinatorial library synthesis: 2001 *J. Comb. Chem.* **4**(5), 369–418
206. Dolle, R. E. (2003) Comprehensive survey of combinatorial library synthesis: 2002 *J. Comb. Chem.* **5**(6), 693–753
207. Dolle, R. E. (2004) Comprehensive survey of combinatorial library synthesis: 2003 *J. Comb. Chem.* **6**(5), 623–679
208. Dolle, R. E. (2005) Comprehensive survey of combinatorial library synthesis: 2004 *J. Comb. Chem.* **7**(6), 739–798

209. Dolle, R. E., Le Bourdonnec, B., Goodman, A. J., Morales, G. A., Salvino, J. M., and Zhang, W. (2007) Comprehensive survey of chemical libraries for drug discovery and chemical biology: 2006 *J. Comb. Chem.* **9** (6), 855–902
210. Dolle, R. E., and Nelson, K. H., Jr. (1999) Comprehensive survey of combinatorial library synthesis: 1998 *J. Comb. Chem.* **1**(4), 235–282
211. Fitzgerald, S. H., Sabat, M., and Geysen, H. M. (2006) Diversity space and its application to library selection and design *J. Chem. Inf. Model.* **46**(4), 1588–1597
212. Rush, T. S., III, Grant, J. A., Mosyak, L., and Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction *J. Med. Chem.* **48**(5), 1489–1495
213. Fitzgerald, S. H., Sabat, M., and Geysen, H. M. (2007) Survey of the diversity space coverage of reported combinatorial libraries *J. Comb. Chem.* **9**(4), 724–734
214. Bergmann, R., Linusson, A., and Zamora, I. (2007) SHOP: Scaffold HOPping by GRID-based similarity searches *J. Med. Chem.* **50**(11), 2708–2717
215. Lead Molecular Design, S.L., Avinguda Cerdanyola 92-94 Local 1.3, 08173 Sant Cugat del Valles, Spain
216. Molecular Discovery Ltd., 215 Marsh Road,1st Floor, HA55NE, Pinner, Middlesex, UK, http://www.moldiscovery.com
217. Goodford, P. J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules *J. Med. Chem.* **28**(7), 849–857
218. Fontaine, F., Pastor, M., Zamora, I., and Sanz, F. (2005) Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent descriptors *J. Med. Chem.* **48**(7), 2687–2694
219. Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004) Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy *J. Med. Chem.* **47**(7), 1739–1749
220. Zamora, I. Private communication
221. Perruccio, F., Mason, J. S., Sciabola, S., and Baroni, M. (2006) FLAP: 4-point pharmacophore fingerprints from GRID *Methods Princ. Med. Chem.* **27**(Molecular Interaction Fields), 83–102
222. Sayle, R., and Nicholls, A. (2006) Electrostatic evaluation of isosteric analogues *J. Comput. Aided Mol. Des.* **20**(4), 191–208
223. Jennings, A., and Tennant, M. (2007) Selection of molecules based on shape and electrostatic similarity: Proof of concept of "Electroforms" *J. Chem. Inf. Model.* **47**(5), 1829–1838
224. Maybridge databases, http://www.maybridge.com/
225. Good, A. C. (2007) Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: Adding flexibility to the search for ligand kin *J. Mol. Graph. Model.* **26**(3), 656–666
226. Cramer, R. D., Cruz, P., Stahl, G., Curtiss, W. C., Campbell, B., Masek, B. B., and Soltanshahi, F. (2008) Virtual screening for R-Groups, including predicted pIC50 contributions, within large structural databases, using Topomer CoMFA *J. Chem. Inf. Model.* **48**(11), 2180–2195
227. Truchon, J. -F., and Bayly, C. I. (2006) Is there a single 'best pool' of commercial reagents to use in combinatorial library design to conform to a desired product-property profile? *Aust. J. Chem.* **59**(12), 879–882
228. Truchon, J. -F., and Bayly, C. I. (2006) GLARE: A new approach for filtering large reagent lists in combinatorial library design using product properties *J. Chem. Inf. Model.* **46**(4), 1536–1548

Chapter 17

The Interweaving of Cheminformatics and HTS

Anne Kümmel and Christian N. Parker

Abstract

The aim of this chapter is to describe the stages of early drug discovery that can be assisted by techniques commonly used in the field of cheminformatics. In fact, cheminformatics tools can be applied all the way from the design of compound libraries and the analysis of HTS results, to the discovery of functional relationships between compounds and their targets.

Key words: High throughput screening, Screening library selection, Hit identification, Hit list triaging, Activity modeling, Multivariate data analysis

1. Introduction

1.1. High Throughput Screening

High throughput screening (HTS) has become one of the main methods for generating leads for drug discovery and for probing basic biological processes [1]. The throughput and range of different targets that can be assessed by screening has increased steadily over the last two decades. This has been in response to advances in human genetics, resulting in many more biological targets being identified and validated [1, 2]. HTS has transformed in vitro biological testing from a process using test tube and cuvette measurements to one using high density, low volume assay formats in automated screening systems. This transformation has been facilitated by the development and utilization of miniaturized, homogeneous assays using novel signal detection methods as well as by the standardization of instrumentation, automation, and also informatics tools and data management systems. These improvements allowed for an increase in reliability and robustness which reduced the personnel workload [3]. Therefore, the screening of hundreds of thousands, and even millions,

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_17, © Springer Science+Business Media, LLC 2011

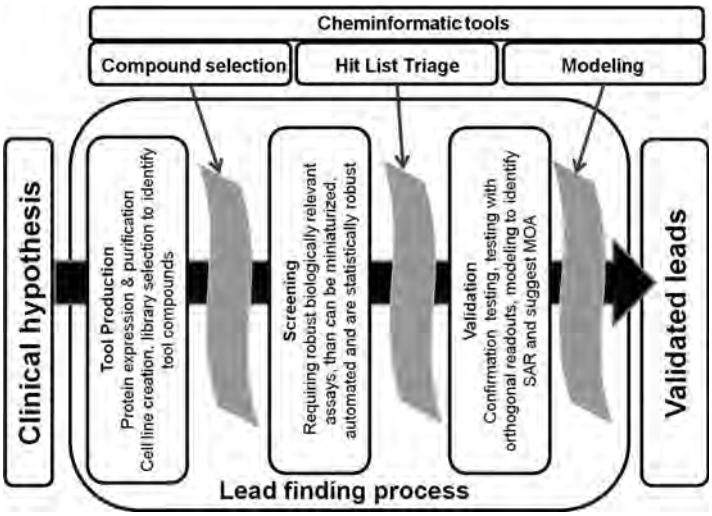


Fig. 1. This figure presents the stages of early drug discovery which must occur before clinical testing can start. The *arrows* indicate the points where computational methods support the process.

of samples has become possible on a regular basis. An overview of the HTS activities involved in the initial phase of drug discovery, i.e. lead identification, is shown in Fig. 1.

Because of the extensive array of different operations involved, HTS should be regarded as a multidisciplinary science. Screening requires activities as diverse as the creation of cell lines, the expression and purification of recombinant proteins by biologists and the design, and synthesis of compound libraries by computational and medicinal chemists. Furthermore, screening includes contributions from engineers, building suitable instruments to allow the automation of biological assays, and from statisticians and modelers who analyze the screening results to extract useful and predictive models of activity. Figure 2 illustrates the interconnection of scientific disciplines required for successful HTS and hit-to-lead optimization processes.

The challenge that HTS scientists face is not only to develop an assay which is biologically relevant and as specific as possible, but that it should also meet the practical constraints of cost, scale and throughput. It is obvious that an improperly designed assay that fails to detect desirable compound classes is useless, even if the screen is trivial or cost effective to execute. However, for assays that are not suitable for HTS and thus cannot be used to test large numbers of compounds it is possible that too little data will be generated to identify promising leads for follow-up. Assays that do not allow for high throughput testing can still be used in a focused screening strategy such as iterative screening or the testing of mixtures of compounds [4]. An HTS campaign must, therefore, be properly designed and configured to maximize screening

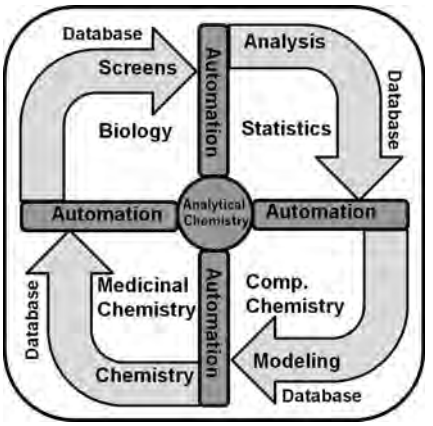


Fig. 2. This figure presents the interconnection of scientific disciplines and activities involved in HTS. Each scientific specialty requires the results from the preceding group and in turn provides the input for the next. Automation includes laboratory robotics as well as data analysis systems and is needed at each step to maximize the throughput of the whole process. Therefore, automation acts as the spokes linking these specialty fields. The central hub of this whole activity is analytical chemistry which is required at every stage of the process to check compound purity and identity as well as to help monitoring biological processes. Finally, all steps need to be connected to a relational database that tracks, records, and integrates all activities.

efficiency as well as ensure biological and pharmacological relevance to be predictive for further lead discovery activities.

Beside the feasible throughput and biological relevance, the readouts that are provided by the screening technology are another parameter that determines the quality of the screening results. Modern screening technologies are increasingly able to monitor multiple assay readouts in order to more comprehensively describe the system being studied. Such assay technologies offer novel opportunities to better describe compound effects and model structure activity relationships. The data analysis and mining of these multivariate measurements is a challenge which requires multivariate data analysis methods, e.g. pattern recognition and machine learning. These methods are already used for the description and organization of chemical compounds using multivariate descriptors. This suggests that there are many opportunities to exploit cheminformatics analysis methods for the analysis of such so-called high content screening assays.

While high throughput screening has become an accepted tool for lead discovery in the pharmaceutical industry, it is also gaining a profound effect on basic biological sciences. This is because HTS is increasingly being applied as a tool to study genetics (i.e. siRNA, antisense, gene overexpression) or pharmacological (small molecule) intervention of biological systems [5, 6]. While siRNA libraries enable the study of the effect of genes on a given biological phenotype, chemical libraries are also

being applied to study biological systems [7]. While small scale screening of chemicals has been used for some time to study biological systems, the advent of the NCBI molecular libraries initiative is having a significant impact, allowing screening to be conducted on a much larger scale than previously possible for academics [8]. This initiative offers a unique opportunity not only to explore the interaction of chemical and biological space but also to rigorously compare different screening strategies and analysis methods, as the results of these screening campaigns are made public. This in turn enables the comparison of different methods to identify optimal screening strategies [9]. The increased availability of such screening data may ultimately provide small molecules interacting with biological targets as starting points to guide rational approaches to find small molecule ligands for any protein [10, 11].

1.2. Overview of Cheminformatics Contribution to HTS Campaigns

Cheminformatics has become a critical tool in the process of high throughput screening and is involved at almost every stage of the process. Prior to screening cheminformatics assists the compound selection (or library design) as well as later on the analysis of HTS screening data, hit triaging and selection. Finally, cheminformatics modeling efforts describe compound effects by structure activity relationship (SAR) and quantitative structure activity relationship (QSAR) models and even predict the mechanisms of action (MOA) for compounds identified in phenotypic screens. Figure 3 presents a flowchart of a generic screening campaign indicating the steps where cheminformatics is involved and in which section of this chapter the corresponding approaches are described.

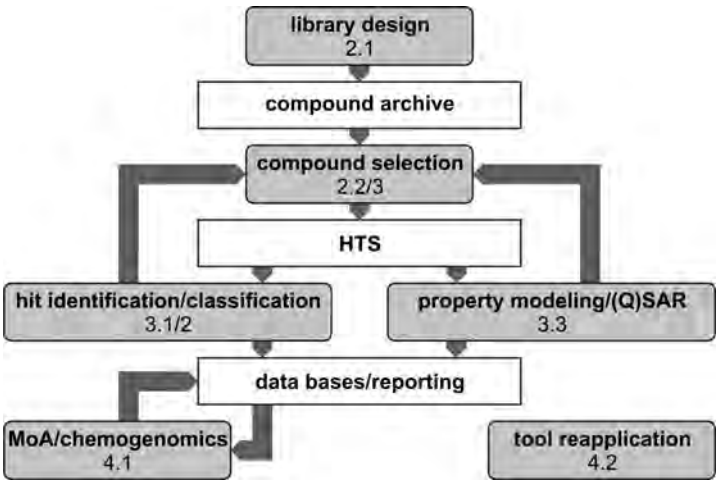


Fig. 3. An overview of the screening process with the steps where cheminformatics are used to support lead identification efforts is presented. The stages in the *grey boxes* are discussed in this review.

2. Assistance in Experimental Design

At the beginning of every HTS campaign the decision which compounds will be tested has to be made. The quality of the final hit list depends on the quality of the library screened and significant effort is thus expended in the design and maintenance of HTS screening libraries [12–14]. Compound selection is critical for screening strategies not only where a limited set of compounds can be tested due to cost or time limitations but also where all available compounds are tested. For the latter, the mere quantity does not necessarily guarantee a satisfactory screening outcome and thus the curation of this collection, e.g. marking of frequent hitters, is necessary to establish and maintain their utility [15–17].

There are a number of different strategies that can be applied when selecting compounds to be screened. One of which is the diversity based approach that aims to select as diverse a range of compounds as possible. Diversity based compound selections have the implicit advantage that such libraries can be used against multiple biological targets and so are often used to select compounds for inclusion in general screening sets [18, 19]. At the other extreme are compound selection strategies that aim to identify a set of compounds for a specific target. Cheminformatics contributes to both the selection of diverse compound libraries (Subsection 2.1) and to the selection of focused sets of compounds chosen for a particular target (Subsection 2.2).

The foundation of many of these strategies is the "similarity concept" which asserts that similar compounds are more likely to have similar activities and properties than dissimilar compounds [20]. This concept was first introduced to cheminformatics in the book "Concepts and Applications of Molecular Similarity" edited by Johnson, M. A. and Maggiora, G. M. which is often quoted but has been out of print for some considerable time [21]. A basic implication of this concept is that the best chance to identifying novel active compounds is to screen as wide a range of dissimilar compounds as possible. Conversely, if a certain compound is known to have the desired activity and properties, compounds having a similar structure should be selected for focused screening.

2.1. Compound Library Design

This section discusses the design of compound libraries that are independent of any particular assay or target. For such a design, general concepts, or filters, are needed that assess the quality and structural diversity of compounds that are included in such a library. These concepts will in consequence also determine the quality and diversity of the hits identified from any screen using these compounds.

As mentioned above, one important concept is that the library should contain as many diverse structures as possible in order to allow for as an unbiased a search for compounds as possible. The goal of such screening strategies is to identify novel leads without missing promising structures. There have been a number of different approaches described for selecting diverse sets of compounds which have been previously reviewed [22]. An interesting approach to assess the coverage of library selection was introduced by Lam and coworkers who used a cell based approach to bin the chemical space of a certain library in lower dimensional spaces [23].

Initially, such diverse libraries were selected from legacy compounds available to pharmaceutical companies but are now more commonly created by carefully purchasing compounds from external vendors [14]. Such a selection should ensure that the library of compounds represents an unbiased set of compounds and is not heavily confined to compound classes previously chosen from medicinal chemistry programs that focused on a particular target. Selecting as diverse a set of compounds from legacy compounds, however, often leads to an accumulation of singletons (single compounds representing a given chemotype) in the lists of active compounds from many screens. This raises the possibility that many chemotypes with active members may have been missed as an insufficient number of compounds from a given scaffold has been tested [24]. In response to this problem, the concept of the chemotype coverage was recently developed. Here, the compounds are assigned 2D-substructures and a sub-library from a reference library is selected such that it covers all 2D-substructures [25].

At first combinatorial chemistry was considered to be a useful source of new compounds to extent the screening collections because of the large numbers of compounds that such combinatorial libraries provided. However, it was quickly realized that the diversity of such libraries was severely limited because of the chemical synthesis methods used to generate such compounds [26, 27]. This has led to a change in focus with combinatorial chemistry increasingly being used to create smaller, focused sets of compounds for hit expansion [28]. However, with the advent of novel chemical synthesis strategies this is again changing as these promise to achieve good coverage of the chemical space [29].

Discussions in the literature as to the correct size of a combinatorial library have led to approaches that estimate the necessary number of compounds from a given chemotype to ensure adequate testing of the respective compound class. Such questions have profound consequences on the number of compounds required for a truly representative diverse screening library. “Back-of-the-envelope” estimates suggest libraries as large as 10–15 million compounds being required to adequately cover

the chemistry space represented by some pharmaceutical companies screening collections [24].

Whilst the initial efforts to design diverse compound libraries were focused on the diversity concept alone, it soon became obvious that for application in the pharmaceutical industry compounds should also be as "drug-like" as possible. Based on the experience of the pharmaceutical research, initially filters and rules to identify toxic or reactive sub-structures were developed, and subsequently the concept of drug-like compounds was established [30, 31]. Cheminformatics models that predict the toxicity, solubility or drug-likeness [32] of chemical structures assist in enriching the library with promising compounds by prioritizing novel compounds or elimination of undesired structures [33].

In addition to the concept of drug-likeness, it has been noted that natural products or their derivatives often represent promising leads [34–36]. Analogous to these concepts, a possible further general criterion to select compounds is their "metabolite-likeness" (i.e. to select compounds that are structurally similar to or have similar properties as metabolites) because metabolites are natural ligands or substrates of pharmacologically relevant targets [37]. As one possibility to assess the similarity of a compound to a specific chemotype or collection a nearest-neighbor approach could be used [38]. This concept was only recently used to compare combinatorial libraries to known drugs and natural products. [39]. It is anticipated that the search for novel areas of chemical space will continue, especially as biologists attempt to address novel classes of biological targets, such as proteins involved in differentiation or development [40, 41].

2.2. Compound/Library Selection Biased Towards Target Classes

There have been a number of reports on the design of compound libraries that focus on specific families of enzymes or receptors. Prominent examples of such families for which targeted libraries were built are kinases, nuclear hormone receptors and G Protein Coupled Receptors (GPCRs). The library design here is based on the known activity information. Compounds that are known to act against proteins/targets of a certain family are used as starting points to select compounds to test against a related target [42–48]. The advantage of such compound libraries is that they often have very high hit rates, making it relatively easy to find tool compounds to further characterize a possible drug target. However, there can be a number of drawbacks with using such focused libraries. For example, it may be difficult to identify compounds with a suitable degree of selectivity against other proteins from the same target class. Another challenge that such focused libraries present is the identification of compounds with sufficient novelty such that intellectual property can be claimed. Because of this difficulty a number of methods to facilitate scaffold hopping have been developed [49, 50]. These methods aim to identify

compounds displaying similar pharmacophore features but not the same compound scaffold.

A limitation of many strategies to design focused compound libraries is that many compounds known to be active against the target or related proteins are needed to build suitable models for library design. In order to circumvent this limitation, the classification of compounds via emerging patterns has recently been developed [51, 52]. This approach is able to select further compounds to screen based on only a few known active compounds. An extension to this strategy to design focused libraries employs natural products as starting points to select for potentially active compounds. The rationale underlying this approach is the notion that the number of structural folds present in proteins is fairly limited and natural products have evolved to interact with these repeated protein folds [10, 53].

Hitherto, the selection of compounds using known ligands for a certain target or target class has been described. However, compounds can also be selected by virtual screening based on target structure information (i.e. docking), even when no ligands are known for the respective target. In cases where the target structure is known docking methods strive to find compounds that would fit into a known (or predicted) binding pocket using a virtual library of compounds. Zhou gives a review on applications of both docking of virtual libraries of compounds or just compound "fragments", which can then be used as starting points for the design of focused compound libraries [29]. If the structure of the target is not known, homology modeling of the target structure using a related target protein can also be applied [54]. To assess the accuracy of such docking procedures many sophisticated methods, including the use of decoy libraries, have been proposed [55, 56]. Nevertheless, the problem remains that these virtual methods have to make assumptions about the activity of a compound against the target. Practical considerations for monitoring the efficiency of docking methods are often limited. Reports usually focus on the discovery of active compounds by screening a small sub-set of compounds that were predicted to bind. Rarely do authors present control experiments showing that an equal number of compounds that are predicted to poorly bind actually fail to bind. Another difficulty that arises when comparing the efficiency of docking methods and other screening methods is that only a small number of docked compounds are chosen for testing. This lower number is usually tested at much higher compound concentrations than usually possible in screens with large numbers of compounds (due to issues of compound consumption, cost and solubility). In consequence, such a screening setup will identify more false positives and promiscuous binders than in an usual HTS [15].

One of the practical difficulties in using focused compound libraries is the need to physically pick and array the chosen

compounds into plates for screening. As a workaround to physically re-arranging compounds and producing novel compound plates, cherry picking or plate selection can be applied [57, 58]. In the approach described in these papers, plates of previously arrayed compounds are selected which represent a maximal diversity of compound chemotypes or molecular features. This approach is similar to a method for optimizing the diversity of combinatorial libraries with a set number of plates to be synthesized as was reported by Agrafiotis and Rassokhin [59]. The plate selection could also analogously be applied for the screening of compounds selected based on a given target. The plate selection process can be modified as new compound plates are added to the library, or when plates of compounds are consumed or retired from a collection.

In the future it is, however, possible that the approaches outlined above will be superseded by modern compound management technologies. New devices enable a flexible compound archive as they are no longer only able to manage compound plates but can rather access individual compounds which may either be stored in plates or vials. These systems would array (or dispense) compounds into screening plates on demand [60]. As this technology enables a more frequent utilization of customized and focused compound sets the need for simple-to-use cheminformatics tools to design these sets will increase.

3. Analysis of HTS Results

3.1. Hit List Generation

Primary screening of large compound selections is usually conducted by testing each compound just once at a single concentration. A number of different hit selection methods have been developed primarily for the selection of hits from such large screening collections.

One of the simplest and most commonly applied methods for selecting hits is to use a simple activity cut-off, i.e. to define a compound as a hit if it presents say at least 50% activity at the concentration screened. However, as has been noted elsewhere, this method can lead to false negatives and thus to a loss of potentially valuable structures due to the inherent error in any biological assay [61, 62].

Other methods for selecting actives include a statistical cut-off to identify compounds that are significantly different from the majority of samples tested. The hit selection threshold is usually set as ± 3 standard deviations (stdev) from the mean of the samples, or in similar manner by using robust measures (median and median absolute deviation) of the distribution of assay readouts. An alternative is to use some deviation from the average of the

inactive controls instead of from all samples to identify compounds altering the assay readout. However, the latter selection criterion is often not appropriate because the controls are usually located in the same place on all compound plates allowing positional effects to bias the recorded assay readout. Additionally, the inactive controls do not contain colored or quenching samples which are often present in compound collections. The hit selection threshold is then often set at too low a value resulting in far too many false positives.

The approach of using the mean ±3 stdev of the samples has the disadvantage that it assumes that the samples being tested are predominantly inactive. This assumption is generally true when large libraries of diverse compounds are being tested, but it is not always accurate if, for example, focused sets of compounds are selected for testing. This assumption also breaks down for agonist screens or screens that detect agonists as well as antagonists. Results from such screens usually do not show a Gaussian distribution and thus the application of a parametric description using mean and stdev (or even median and MAD) of the results is questionable.

The cut-off for hit selection can also be adapted to each plate by selecting the top 1% (or so) of compounds showing activity on a single plate [62]. For screening formats using higher plate densities (such as 384 and 1,536) this works very well and ensures that the "best" samples from each plate are selected. The plate-specific adaption of the cut-off will allow weak hits to be detected on plates with only few active compounds. There is, however, the risk to miss out true actives from plates containing many active compounds as can happen with plates generated using combinatorial chemistry or focused compound libraries. This approach is very similar to the methods evolving for hit selection from screening biological samples such as siRNA [63]. More sophisticated modeling approaches have also been described, including methods to model the activity distribution of active and inactive compounds in the results [61]. However, this approach has not become widely accepted possibly due to the time and computational expertise needed to conduct such an analysis.

Usually, hit selection does not involve cheminformatics analysis of the screening readout. There are, however, examples that demonstrate the potential of modeling efforts to improve hit selection by taking structural information of the tested compounds into account [64]. These approaches often aim to rescue compounds that have fallen below the hit selection cut-off and thus they still depend on setting an initial activity cut-off. In essence, the model is generated using a subset, or even all, of the available screening data and is then used to predict the activity of the tested compounds. Those compounds, predicted to be active but failing to show activity in the screen, can then be considered

as false negatives and can be retested. Another approach is to use the information about active compounds to generate a phylogenic tree organizing the compounds by common substructures. This hierarchical clustering can then be used to mine the inactive compounds for possible false negatives or SARs [65]. As an alternative approach to hit selection the compounds are classified based on structure to then identify active compound classes instead of single compounds. These strategies are discussed for example by MacFayden et al., who were able to show enrichment of actives for compounds from a given structure family, [66]. Another approach to prioritizing compounds classes rather than individual compounds for further evaluation uses similarity searching around both active and inactive compounds [33, 67]. Also, a rigorous statistical evaluation of the data using structural information about the compounds was suggested to base hit identification on the probability that a specific hit comes from a given structure family [68]. This approach accounts for the prevalence of any given number of compounds in any given chemotype. The confirmation rate of compounds chosen as possible actives using this method was higher compared to traditional hit identification based on a global activity threshold.

Cheminformatics tools will increasingly support the hit selection process, especially as there has been a movement away from screening at single concentrations to monitoring assay responses over a range of concentration (i.e. testing multiple compound concentrations) [69]. This approach has the advantage that the primary screening data immediately provides results suitable for QSAR analysis. However, there has been no in-depth evaluation of the cost effectiveness of this approach. As concentration response screening (also known as qHTS) requires testing multiple concentrations of the same compound and as most compounds tested in a screen are inactive, such a strategy could place significant additional costs on screening. This caveat is especially true if a diverse set of compounds is being tested as the number of active compounds from any one structural class will be limited. Thus, the extraction of any structure activity relationships for a class of active compounds is restricted, if not impossible. In comparison to a qHTS, screening at a single concentration and expanding on the identified active compounds may still be a more cost effective strategy. Further comparison and evaluation of the effectiveness of these strategies still needs to be assessed.

All of the methods described above have been developed using a single assay readout to describe the compound activity. However, assays to monitor biological systems are increasingly able to monitor multiple parameters to describe the effect of a compound on the system. This then creates an issue of how best to select hits using these multiparametric data sets. This is yet another opportunity where cheminformatics could help to improve the

drug discovery process. Multivariate methods of classification and selection are well known and applied by computational chemists. This knowledge can readily be reapplied to analysis of multivariate readouts from high content screening and we will discuss this further in Subsection 4.3.

3.2. Triage of Hit Lists

Similar to removing compounds that are reactive or have undesired moieties for library design as mentioned in Subsection 2.1, cheminformatics tools are used to triage such compounds from the active compounds as identified in a HTS. This has historically been necessary because compound archives have been acquired over time with screening libraries including reactive intermediates and other undesired features. But as described in Subsection 2.1 significant effort has been extended in cleaning up screening collections to remove such compounds before they are ever tested as reported by Verheij [31].

However, cheminformatics tools are still needed to remove compounds that may have undesired biological effects or that are frequent hitters as predicted for example with Bayesian modeling [15, 70]. As a mean to assess whether a compound may have adverse off-target effects a modeling approach was proposed that is able to predict the activity of a compound against multiple possible targets [71]. Other approaches that are applied in special cases triage or prioritize compounds on the basis whether they presumably display the desired pharmaco-distribution within the organism. For example, compounds are selected that are (or are not) able to cross the blood brain barrier [72]. These more specialized models that are focused on selecting more defined characteristics seem to be gaining more acceptance than general methods that triage compounds based on their potential drug-like characteristics [73].

In addition to hit list triaging it is also possible to identify, or at least suggest, possible false negatives or positives before continuing the evaluation of the hit list. A number of different methods are available to relate the structure of the compounds to the observed biological activity. A comparison of different machine learning techniques (Bayesian Classifier, support vector machine and recursive partitioning) demonstrated their robustness when predicting compound activity in order to suggest false negatives [74].

3.3. Modeling and Classification of Active Compounds

There is a wide range of cheminformatics approaches to explore the (validated) hit lists further. One goal here is to suggest further available compounds to be tested. Another important goal is to understand the structure–activity relationship (SAR) that the respective compound classes possess. This may guide the synthesis of novel compounds for testing based on the structural features contributing to the desired activity that were revealed based on

the SAR model. The methods applied to explore the screening results often involve grouping of the structures that were tested which can be done in two basic ways, clustering or classification. As mentioned in Subsection 3.1 such organization of hits can also be used as a means of identifying compounds that may be false negatives or even false positives.

As noted above, possibly the most widely used approach for hit expansion is to search for structurally similar compounds to the identified active compounds [75]. A wide range of different methods exist that make use of different molecular descriptors or similarity measures [20]. These different setups have been shown to predict different sets of similar compounds for a particular hit. Therefore, the joint application of several methods can help to explore the SAR landscape around a particular hit [76]. To more rationally exploit such results, data fusion methods have been suggested to efficiently capture the different types of compounds that can be identified [77]. Although not explicitly developed for hit expansion methods to visualize and describe, SAR landscapes have been presented [78]. Other methods to efficiently explore around hits from screening include approaches that seek to use limited information about active compounds to identify additional active compounds with similar physico-chemical properties [79]. Methods which help to visualize the SAR around a hit also include the Scaffold tree concept which can be used to guide the discovery of additional active compounds that may have not already been tested or even available at the time of primary screening [80].

A number of different ways for clustering compounds have been described using different clustering algorithms and molecular descriptors [81]. However, it should be noted that while clustering of compounds is very useful and can provide an unbiased way of exploring possible SARs there are a number of important drawbacks. The first drawback is that the clustering to a large extent depends on the set of compounds being used. The addition of a new compound or the removal of a compound from the set can generate a different clustering in which compounds that previously belonged to the same group may be assigned to different groups afterwards. In addition the use of different descriptors or similarity measures can again result in different clustering of the compounds. This instability can be confusing and makes comparison of different sets of results difficult. (Nevertheless, the use of different descriptors or clustering methods offers the opportunity to explore and identify additional relationships between compounds.) Clustering methods have been extended to allow maximum common substructures to be identified within compound clusters [82]. While such methods still have many of the drawbacks common to clustering methods here a core or common scaffold for each cluster can be identified which can then be used

to search the compound library using substructure searches rather than similarity searches.

A strategy to avoid the problems associated with compound clustering is to use classification schemes in order to organize compounds. Such classification schemes are based on rules how to describe the compound structures with prototype substructures such as scaffolds and side chains. One of the first examples of compound classification to receive regular use were the so-called Murcko scaffolds [83]. However, this approach has been expanded to look at general molecular scaffolds of varying specificity, e.g. Meqnims or SCINS classification schemes [84]. These hierarchical classification schemes were further developed to allow even complex natural products to be classified in an ordered and hierarchical manner [85]. The advantage of these classification schemes is that these organizations are consistent and can be applied to compounds in a pre-determined manner and often allow the compounds to be classified into groups of increasing molecular specificity.

As an application not only to describe and organize compounds based on their structure it has been suggested that such a compound ontology could be used to predict the function or activity of a compound in a cellular environment [86]. Subsequently, efforts have been made to link such compound classifications with gene ontologies [87]. Such strategies of using screening information from groups of related compounds to better estimate the activity of a given compound class can also be applied to screening siRNA libraries where gene set enrichment can be used to improve the selection of possible active siRNAs, e.g. [88].

4. Advances and Emerging Techniques

4.1. Integration of Biological Activity Information with Cheminformatics Approaches: Chemogenomics

The integration of cheminformatics and biology results has been extensively reviewed [89, 90]. The main aim of such analysis has been to find correlations between the chemical structure of a compound and the profile of biological activities [91, 92]. Such analyses are being extended to allow the prediction of off-target effects [93]. And even to explore the possibility to predict complex biological readouts such as unwanted side effects in clinical trials [94, 95]. The challenge facing such efforts is, as noted previously, that biological assays are increasingly monitoring multiple parameters and thus are more complex to be analyzed. Another trend that has begun to receive attention is the use of screening to monitor the effect of compound combinations, looking for synergies or to help delineate possible mechanism of action [96]. Here again new methods to correlate the structure of the compounds in the mixture with the observed biological effect will be needed.

4.2. Re-application of Cheminformatics Tools to Multivariate HTS Results

Modern screening technologies (e.g. HCS or FACS) are increasingly providing multivariate readouts. So it can be foreseen that better profiling of a compounds' effect on an assay can be achieved, rather than only classifying compounds into hits and non-hits based on a single readout (e.g. [97, 98]). In addition screening results from different assays could also be brought together for a detailed profile of the compound activity (e.g. toxicity measurements or pharmacodynamic parameters). However, the statistical and mathematical background for the in-depth data mining of such screening results is not a skill readily available to many screeners. As cheminformatics groups and screening labs are already working closely together on screening projects and as the former know about multivariate data analysis there is a potential to develop methods for the analysis of multivariate assays by drawing on the skills of computational chemists.

Before applying a method that is already established in cheminformatics changes may have to be made to adapt the methods to be suitable for the screening results. For example, the Tanimoto distance that is applied for comparing binary fingerprints might be substituted by the Euclidean distance for analyzing continuous readouts. Or vice versa, the continuous readouts of biological assays could be converted to be binary or integers such that methods developed in cheminformatics such as median partitioning can be applied [99]. The idea of biological fingerprints which utilizes multivariate experimental readouts has already been introduced to the field of cheminformatics as an additional compound descriptor [93].

One requirement is the reduction of the dimensionality space obtained from screening assays which can be in the hundreds for image based screening technology [100, 101]. Different dimension reduction and feature selection methods have been compared in the field of HCS [102]. Factor analysis was used to derive a reduced set of biologically interpretable parameters [103]. Principle component analysis has also proved to be useful for visualization and exploring of HCA results [104]. Dimension reduction is also applied in cheminformatics to either reduce the dimensionality of the chemical descriptor [105, 106] or to emphasize which molecular features descriptors achieve better prediction of hit (e.g. [107, 108]). Similar procedures would be useful in assay development to determine the importance of the image readouts to discriminate control compounds. This information could be used to better understand the effect of the tool compounds to the assay and to select the set of monitored readouts during the high throughput screening.

Hit identification based on multiple parameters can draw on clustering and classification techniques which are already established and used in cheminformatics. Hit selection and compound profiling using HCS has been conducted using a number of data

mining methods such as k-nearest neighbors (KNN), support vector machines (SVM), linear discriminant analysis (LDA) [109–111] which have also been applied to cheminformatic data analysis as well (a comparison of different method is described in [112]). Either unsupervised or supervised clustering/classification could be applied for hit or compound profiling. For the latter, control compounds have to be available that are used to train the classifier. SVM is an example that has already been applied to both HCS hit identification [109, 110, 113] and cheminformatics [74, 114]. Recently, another approach has been taken to summarize multidimensional data sets to a single, relevant, summary value using LDA [115]. This method has been used to characterize assay performance but has not yet been used for hit selection which is an obvious additional application. Predictive models based on Bayesian modeling or partition trees that are widely used in cheminformatics have not yet been applied to multivariate screening results. A bottleneck for the application of such methods is the necessity of having tool compounds that are not always available. Also, in contrast to chemical descriptors, the multivariate response characterization is assay specific and thus general models cannot be used in a screening campaign unless general assays such as cytotoxicity profiling are performed.

The mere exploration of the screening results by unsupervised clustering could help to prioritize groups of compounds for further testing and lead optimization. For example, hierarchical clustering was used to organize compounds based on their multiple readout profile to explore off-target effects and hidden phenotypes [116].

Another field, in which unsupervised clustering techniques could be applied, is the detection and analysis of cellular subpopulations. In one example for a HCS assay it was possible to identify subpopulations of cells using KNN clustering [111] which also is frequently used in cheminformatics to determine similar sets of chemical structure (e.g. [117]). The identification of subpopulations is routinely conducted for fluorescent activated cell sorting (FACS). Until today, the identification of subpopulations is usually accomplished by manual assessment of scatter plots. However, automatic subpopulation analyses are being developed not least trigged by improvements allowing for a larger number of fluorescent readouts to be monitored with this technique [118, 119].

5. Conclusion and Discussion

The aim of this chapter has been to show how every step of a screening campaign – from compound selection to analyzing

and visualizing the final results – benefits from the support of cheminformatics.

Possibly one of the most important questions facing cheminformatics and high throughput screening is what is the optimal size of a library to screen? There are many practical constraints that any screening group is faced with (be they as simple as compound availability, available space to store compounds or the constraints of time or cost in running the biological assay of interest) and which all make it important to find ways to rationally optimize the number of compounds tested as well as to rationally select compounds. This involves also questions like whether it is better to test compounds multiple times at the same concentration or to test them with a concentration response manner? Or, how many compounds from each compound class should be tested? For this reason direct experimental evidence comparing different screening strategies needs to be reported and challenged within the scientific literature, so that researchers are able to choose optimal solutions. For example, a detailed comparison of cost efficiency of screening using a single concentration versus multiple concentrations has yet to be made [69]. This comparison would trade off the added cost of screening with multiple doses against the possibility to miss active compounds by screening at a single concentration.

The use of focused compound libraries (or known pharmacologically active compounds and the library used for the SOSA approach [120]) also needs to be rigorously evaluated and compared to screening of new diverse areas of chemistry. While such focused strategies clearly have the advantage of allowing active compounds to be identified the challenges of identifying patentable compounds with sufficient selectivity has to be compared to the difficulties of identifying truly novel compounds.

One area where the application of rational screening strategies is lacking, but strongly needed, is to decide on the number of compounds tested when screening natural product extracts. The reason is that at present there is no quantitative way to describe or organize the different extracts generated from natural sources. As a result screeners are left with having to simply screen as many extracts as are available to them. Strategies to organize the samples by the species, genus or class producing the extracts are not suitable as even the same species can generate different metabolites depending on the media they are grown in. However, with the increasing availability of metabolomic profiling it may be possible in the near future to profile natural product extracts. Such fingerprints that characterize the extracts would enable the application of cheminformatics methods to classify and prioritize certain extracts.

References

1. Mayr, L. M., and Fuerst, P. (2008) The future of high-throughput screening. *J Biomol Screen* **13**, 443–448.
2. Macarron, R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov Today* **11**, 277–279.
3. Pereira, D. A., and Williams, J. A. (2007) Origin and evolution of high throughput screening. *Br J Pharmacol* **152**, 53–61.
4. Ferrand, S., Schmid, A., Engeloch, C., and Glickman, J. F. (2005) Statistical evaluation of a self-deconvoluting matrix strategy for high-throughput screening of the CXCR3 receptor. *Assay Drug Dev Technol* **3**, 413–424.
5. Gaither, L. A. (2007) Chemogenomics approaches to novel target discovery. *Expert Rev Proteomics* **4**, 411–419.
6. Jacoby, E., Schuffenhauer, A., and Floersheim, P. (2003) Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect* **16**, 93–102.
7. Ding, L., Paszkowski-Rogacz, M., Nitzsche, A., Slabicki, M. M., Heninger, A. K., de Vries, I., Kittler, R., Junqueira, M., Shevchenko, A., Schulz, H., Hubner, N., Doss, M. X., Sachinidis, A., Hescheler, J., Iacone, R., Anastassiadis, K., Stewart, A. F., Pisabarro, M. T., Caldarelli, A., Poser, I., Theis, M., and Buchholz, F. (2009) A genome-scale RNAi screen for Oct4 modulators defines a role of the Paf1 complex for embryonic stem cell identity. *Cell Stem Cell* **4**, 403–415.
8. Inglese, J., Johnson, R. L., Simeonov, A., Xia, M., Zheng, W., Austin, C. P., and Auld, D. S. (2007) High-throughput screening assays for the identification of chemical probes. *Nat Chem Biol* **3**, 466–479.
9. Oprea, T. I., Bologa, C. G., Boyer, S., Curpan, R. F., Glen, R. C., Hopkins, A. L., Lipinski, C. A., Marshall, G. R., Martin, Y. C., Ostopovici-Halip, L., Rishton, G., Ursu, O., Vaz, R. J., Waller, C., Waldmann, H., and Sklar, L. A. (2009) A crowdsourcing evaluation of the NIH chemical probes. *Nat Chem Biol* **5**, 441–447.
10. Dekker, F. J., Koch, M. A., and Waldmann, H. (2005) Protein structure similarity clustering (PSSC) and natural product structure as inspiration sources for drug development and chemical genomics. *Curr Opin Chem Biol* **9**, 232–239.
11. Schreiber, S. L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**, 1964–1969.
12. Bologa, C. G., Olah, M. M., and Oprea, T. I. (2006) Chemical database preparation for compound acquisition or virtual screening. *Methods Mol Biol* **316**, 375–388.
13. Olah, M. M., Bologa, C. G., and Oprea, T. I. (2004) Strategies for compound selection. *Curr Drug Discov Technol* **1**, 211–220.
14. Schuffenhauer, A., Popov, M., Schopfer, U., Acklin, P., Stanek, J., and Jacoby, E. (2004) Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb Chem High Throughput Screen* **7**, 771–781.
15. Crisman, T. J., Parker, C. N., Jenkins, J. L., Scheiber, J., Thoma, M., Kang, Z. B., Kim, R., Bender, A., Nettles, J. H., Davies, J. W., and Glick, M. (2007) Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J Chem Inf Model* **47**, 1319–1327.
16. McGovern, S. L., Helfand, B. T., Feng, B., and Shoichet, B. K. (2003) A specific mechanism of nonspecific inhibition. *J Med Chem* **46**, 4265–4272.
17. Seidler, J., McGovern, S. L., Doman, T. N., and Shoichet, B. K. (2003) Identification and prediction of promiscuous aggregating inhibitors among known drugs. *J Med Chem* **46**, 4477–4486.
18. Austin, C. P., Brady, L. S., Insel, T. R., and Collins, F. S. (2004) NIH molecular libraries initiative. *Science* **306**, 1138–1139.
19. Lajiness, M., and Watson, I. (2008) Dissimilarity-based approaches to compound acquisition. *Curr Opin Chem Biol* **12**, 366–371.
20. Martin, Y. C. (2001) Diverse viewpoints on computational aspects of molecular diversity. *J Comb Chem* **3**, 231–250.
21. Johnson, M. A., and Maggiora, G. M. (1990) Concepts and Application of Molecular Similarity. John Wiley & Sons: New York.
22. Xue L. (2000) Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb Chem High Throughput Screen* **3**, 363–372.
23. Lam, R. L. H., Welch, W. J., and Young, S. S. (2002) Uniform coverage designs for molecule selection. *Technometrics* **44**, 99–109.
24. Lipkin, M. J., Stevens, A. P., Livingstone, D. J., and Harris, C. J. (2008) How large does a compound screening collection need to be? *Comb Chem High Throughput Screen* **11**, 482–493.

25. Johnson, M., Shanmugasundaram, V., Bundy, G., Chapman, D., and Kilkuskie, R. (2009) Chemotypic coverage: a new basis for constructing screening sublibraries. *J Chem Inf Model* **49**, 531–542.
26. Fitzgerald, S. H., Sabat, M., and Geysen, H. M. (2007) Survey of the diversity space coverage of reported combinatorial libraries. *J Comb Chem* **9**, 724–734.
27. Krier, M., Bret, G., and Rognan, D. (2006) Assessing the scaffold diversity of screening libraries. *J Chem Inf Model* **46**, 512–524.
28. Fotouhi, N., Gillespie, P., Goodnow, R. A., So, S. S., Han, Y., and Babiss, L. E. (2006) Application and utilization of chemoinformatics tools in lead generation and optimization. *Comb Chem High Throughput Screen* **9**, 95–102.
29. Zhou, J. Z. (2008) Structure-directed combinatorial library design. *Curr Opin Chem Biol* **12**, 379–385.
30. Rishton, G. M. (1997) Reactive compounds and in vitro false positives in HTS. *Drug Discov Today* **2**, 382–384.
31. Verheij, H. J. (2006) Leadlikeness and structural diversity of synthetic screening libraries. *Mol Divers* **10**, 377–388.
32. Vistoli, G., Pedretti, A., and Testa, B. (2008) Assessing drug-likeness – what are we missing? *Drug Discov Today* **13**, 285–294.
33. Schreyer, S. K., Parker, C. N., and Maggiora, G. M. (2004) Data shaving: a focused screening approach. *J Chem Inf Comput Sci* **44**, 470–479.
34. Ertl, P., Roggo, S., and Schuffenhauer, A. (2008) Natural product-likeness score and its application for prioritization of compound libraries. *J Chem Inf Model* **48**, 68–74.
35. Grabowski, K., Baringhaus, K. H., and Schneider, G. (2008) Scaffold diversity of natural products: inspiration for combinatorial library design. *Nat Prod Rep* **25**, 892–904.
36. Kaiser, M., Wetzel, S., Kumar, K., and Waldmann, H. (2008) Biology-inspired synthesis of compound libraries. *Cell Mol Life Sci* **65**, 1186–1201.
37. Dobson, P. D., Patel, Y., and Kell, D. B. (2009) 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discov Today* **14**, 31–40.
38. Guha, R., Dutta, D., Jurs, P. C., and Chen, T. (2006) R-NN curves: an intuitive approach to outlier detection using a distance based method. *J Chem Inf Model* **46**, 1713–1722.
39. Singh, N., Guha, R., Giulianotti, M. A., Pinilla, C., Houghten, R. A., and Medina-Franco, J. L. (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J Chem Inf Model* **49**, 1010–1024.
40. Borowiak, M., Maehr, R., Chen, S., Chen, A. E., Tang, W., Fox, J. L., Schreiber, S. L., and Melton, D. A. (2009) Small molecules efficiently direct endodermal differentiation of mouse and human embryonic stem cells. *Cell Stem Cell* **4**, 348–358.
41. Lyssiotis, C. A., Foreman, R. K., Staerk, J., Garcia, M., Mathur, D., Markoulaki, S., Hanna, J., Lairson, L. L., Charette, B. D., Bouchez, L. C., Bollong, M., Kunick, C., Brinker, A., Cho, C. Y., Schultz, P. G., and Jaenisch, R. (2009) Reprogramming of murine fibroblasts to induced pluripotent stem cells with chemical complementation of Klf4. *Proc Natl Acad Sci USA* **106**, 8912–8917.
42. Akritopoulou-Zanze, I., and Hajduk, P. J. (2009) Kinase-targeted libraries: the design and synthesis of novel, potent, and selective kinase inhibitors. *Drug Discov Today* **14**, 291–297.
43. Bocker, A., Sasse, B. C., Nietert, M., Stark, H., and Schneider, G. (2007) GPCR targeted library design: novel dopamine D3 receptor ligands. *ChemMedChem* **2**, 1000–1005.
44. Deanda, F., Stewart, E. L., Reno, M. J., and Drewry, D. H. (2008) Kinase-targeted library design through the application of the PharmPrint methodology. *J Chem Inf Model* **48**, 2395–2403.
45. Guo, T., and Hobbs, D. W. (2003) Privileged structure-based combinatorial libraries targeting G protein-coupled receptors. *Assay Drug Dev Technol* **1**, 579–592.
46. Hoppe, C., Steinbeck, C., and Wohlfahrt, G. (2006) Classification and comparison of ligand-binding sites derived from grid-mapped knowledge-based potentials. *J Mol Graph Model* **24**, 328–340.
47. Mestres, J., Martin-Couce, L., Gregori-Puigjane, E., Cases, M., and Boyer, S. (2006) Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J Chem Inf Model* **46**, 2725–2736.
48. Shuttleworth, S. J., Connors, R. V., Fu, J., Liu, J., Lizarzaburu, M. E., Qiu, W., Sharma, R., Wanska, M., and Zhang, A. J. (2005) Design and synthesis of protein superfamily-targeted chemical libraries for lead identification and optimization. *Curr Med Chem* **12**, 1239–1281.
49. Cheeseright, T. J., Holm, M., Lehmann, F., Luik, S., Gottert, M., Melville, J. L., and Laufer, S. (2009) Novel lead structures for

p38 MAP kinase via FieldScreen virtual screening. *J Med Chem* **52**, 4200–4209.
50. Jenkins, J. L., Glick, M., and Davies, J. W. (2004) A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J Med Chem* **47**, 6144–6159.
51. Auer, J., and Bajorath, J. (2006) Emerging chemical patterns: a new methodology for molecular classification and compound selection. *J Chem Inf Model* **46**, 2502–2514.
52. Auer, J., and Bajorath, J. (2008) Simulation of sequential screening experiments using emerging chemical patterns. *Med Chem* **4**, 80–90.
53. Koch, M. A., Wittenberg, L. O., Basu, S., Jeyaraj, D. A., Gourzoulidou, E., Reinecke, K., Odermatt, A., and Waldmann, H. (2004) Compound library development guided by protein structure similarity clustering and natural product structure. *Proc Natl Acad Sci USA* **101**, 16721–16726.
54. Rockey, W. M., and Elcock, A. H. (2006) Structure selection for protein kinase docking and virtual screening: homology models or crystal structures? *Curr Protein Pept Sci* **7**, 437–457.
55. Huang, N., Shoichet, B. K., and Irwin, J. J. (2006) Benchmarking sets for molecular docking. *J Med Chem* **49**, 6789–6801.
56. von, Korff. M., Freyss, J., and Sander, T. (2009) Comparison of ligand and structure-based virtual screening on the DUD data set. *J Chem Inf Model* **49**, 209–231.
57. Crisman, T. J., Jenkins, J. L., Parker, C. N., Hill, W. A., Bender, A., Deng, Z., Nettles, J. H., Davies, J. W., and Glick, M. (2007) "Plate cherry picking": a novel semi-sequential screening paradigm for cheaper, faster, information-rich compound selection. *J Biomol Screen* **12**, 320–327.
58. Sukuru, S. C., Jenkins, J. L., Beckwith, R. E., Scheiber, J., Bender, A., Mikhailov, D., Davies, J. W., and Glick, M. (2009) Plate-based diversity selection based on empirical hts data to enhance the number of hits and their chemical diversity. *J Biomol Screen* **14**, 690–699.
59. Agrafiotis, D. K., and Rassokhin, D. N. (2001) Design and prioritization of plates for high-throughput screening. *J Chem Inf Comput Sci* **41**, 798–805.
60. Andreae, M. R., Steiner, T., Hueber, M., Schopfer, U., Smith, R., Igo, D., Cantrell, D., Hohos, A., and Kiwanuka, A. (2008) Closing the gap between centralized and decentralized compound management approaches. *Comb Chem High Throughput Screen* **11**, 825–833.
61. Buxser, S., and Chapman, D. L. (2007) Use of mixture distributions to deconvolute the behavior of "hits" and controls in high-throughput screening data. *Anal Biochem* **361**, 197–209.
62. Gubler, H. (2006) In: High-Throughput Screening in Drug Discovery Hüser, J. (Ed.) Handling and Management of Primary Assay Data, Wiley-VCH: Weinheim.
63. Birmingham, A., Selfors, L. M., Forster, T., Wrobel, D., Kennedy, C. J., Shanks, E., Santoyo-Lopez, J., Dunican, D. J., Long, A., Kelleher, D., Smith, Q., Beijersbergen, R. L., Ghazal, P., and Shamu, C. E. (2009) Statistical methods for analysis of high-throughput RNA interference screens. *Nat Methods* **6**, 569–575.
64. Cloutier, L. M., and Sirois, S. (2008) Bayesian versus Frequentist statistical modeling: a debate for hit selection from HTS campaigns. *Drug Discov Today* **13**, 536–542.
65. Nicolaou, C. A., Tamura, S. Y., Kelley, B. P., Bassett, S. I., and Nutt, R. F. (2002) Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J Chem Inf Comput Sci* **42**, 1069–1079.
66. MacFayden, I., Walker, G., and Alvarez, J. (2005) In: Chemoinformatics in Drug Discovery Oprea, T. I. (Ed.) Enhancing hit quality and diversity within assay throughput constraints, Wiley-VCH: Weinheim, pp 143–173.
67. Yan, S. F., King, F. J., He, Y., Caldwell, J. S., and Zhou, Y. (2006) Learning from the data: mining of large high-throughput screening databases. *J Chem Inf Model* **46**, 2381–2395.
68. Yan, S. F., Asatryan, H., Li, J., and Zhou, Y. (2005) Novel statistical approach for primary high-throughput screening hit selection. *J Chem Inf Model* **45**, 1784–1790.
69. Inglese, J., Auld, D. S., Jadhav, A., Johnson, R. L., Simeonov, A., Yasgar, A., Zheng, W., and Austin, C. P. (2006) Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc Natl Acad Sci USA* **103**, 11473–11478.
70. Azzaoui, K., Hamon, J., Faller, B., Whitebread, S., Jacoby, E., Bender, A., Jenkins, J. L., and Urban, L. (2007) Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2**, 874–880.
71. Bender, A., Jenkins, J. L., Glick, M., Deng, Z., Nettles, J. H., and Davies, J. W. (2006) "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are

multitarget drugs a feasible concept? *J Chem Inf Model* **46**, 2445–2456.
72. Kortagere, S., Chekmarev, D., Welsh, W. J., and Ekins, S. (2008) New predictive models for blood-brain barrier permeability of drug-like molecules. *Pharm Res* **25**, 1836–1845.
73. Blake, J. F. (2000) Chemoinformatics – predicting the physicochemical properties of 'drug-like' molecules. *Curr Opin Biotechnol* **11**, 104–107.
74. Glick, M., Jenkins, J. L., Nettles, J. H., Hitchings, H., and Davies, J. W. (2006) Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J Chem Inf Model* **46**, 193–200.
75. Shanmugasundaram, V., Maggiora, G. M., and Lajiness, M. S. (2005) Hit-directed nearest-neighbor searching. *J Med Chem* **48**, 240–248.
76. Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., and Van Drie, J. H. (2009) Navigating structure–activity landscapes. *Drug Discov Today* **14**, 698–705.
77. Whittle, M., Gillet, V. J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: similarity and group fusion. *J Chem Inf Model* **46**, 2206–2219.
78. Guha, R., and Van Drie, J. H. (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. *J Chem Inf Model* **48**, 646–658.
79. Lounkine, E., Hu, Y., Batista, J., and Bajorath, J. (2009) Relevance of feature combinations for similarity searching using general or activity class-directed molecular fingerprints. *J Chem Inf Model* **49**, 561–570.
80. Renner, S., van Otterlo, W. A., Dominguez, S. M., Mocklinghoff, S., Hofmann, B., Wetzel, S., Schuffenhauer, A., Ertl, P., Oprea, T. I., Steinhilber, D., Brunsveld, L., Rauh, D., and Waldmann, H. (2009) Bioactivity-guided mapping and navigation of chemical space. *Nat Chem Biol* **5**, 585–592.
81. Maggiora, G. M., and Shanmugasundaram, V. (2004) Molecular similarity measures. *Methods Mol Biol* **275**, 1–50.
82. Tamura, S. Y., Bacha, P. A., Gruver, H. S., and Nutt, R. F. (2002) Data analysis of high-throughput screening results: application of multidomain clustering to the NCI anti-HIV data set. *J Med Chem* **45**, 3082–3093.
83. Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **39**, 2887–2893.
84. Xu, Y. J., and Johnson, M. (2002) Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J Chem Inf Comput Sci* **42**, 912–926.
85. Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M. A., and Waldmann, H. (2007) The scaffold tree – visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model* **47**, 47–58.
86. McShan, D. C. (2005) Towards inference of a biochemical ontology from a metabolic database. *Comp Funct Genomics* **6**, 398–406.
87. Yang, J. O., Charny, P., Lee, B., Kim, S., Bhak, J., and Woo, H. G. (2007) GS2PATH: a web-based integrated analysis tool for finding functional relationships using gene ontology and biochemical pathway data. *Bioinformation* **2**, 194–196.
88. Bankhead, A., III, Sach, I., Ni, C., LeMeur, N., Kruger, M., Ferrer, M., Gentleman, R., and Rohl, C. (2009) Knowledge based identification of essential signaling from genome-scale siRNA experiments. *BMC Syst Biol* **3**, 80.
89. Jacoby, E., Bouhelal, R., Gerspacher, M., and Seuwen, K. (2006) The 7 TM G-protein-coupled receptor target family. *ChemMedChem* **1**, 761–782.
90. Marechal, E. (2008) Chemogenomics: a discipline at the crossroad of high throughput technologies, biomarker research, combinatorial chemistry, genomics, cheminformatics, bioinformatics and artificial intelligence. *Comb Chem High Throughput Screen* **11**, 583–586.
91. Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005) Biological spectra analysis: linking biological activity profiles to molecular structure. *Proc Natl Acad Sci USA* **102**, 261–266.
92. Fliri, A. F., Loging, W. T., Thadeio, P. F., and Volkmann, R. A. (2005) Biospectra analysis: model proteome characterizations for linking molecular structure and biological response. *J Med Chem* **48**, 6918–6925.
93. Bender, A., Young, D. W., Jenkins, J. L., Serrano, M., Mikhailov, D., Clemons, P. A., and Davies, J. W. (2007) Chemogenomic data analysis: prediction of small-molecule targets and the advent of biological fingerprint. *Comb Chem High Throughput Screen* **10**, 719–731.
94. Scheiber, J., Chen, B., Milik, M., Sukuru, S. C., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., Glick, M., Davies, J. W., and Jenkins, J. L. (2009) Gaining insight into off-target mediated

effects of drug candidates with a comprehensive systems chemical biology analysis. *J Chem Inf Model* **49**, 308–317.
95. Scheiber, J., Jenkins, J. L., Sukuru, S. C., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J., Urban, L., Glick, M., and Davies, J. W. (2009) Mapping adverse drug reactions in chemical space. *J Med Chem* **52**, 3103–3107.
96. Lehar, J., Krueger, A. S., Avery, W., Heilbut, A. M., Johansen, L. M., Price, E. R., Rickles, R. J., Short, G. F., III, Staunton, J. E., Jin, X., Lee, M. S., Zimmermann, G. R., and Borisy, A. A. (2009) Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat Biotechnol* **27**, 659–666.
97. Adams, C. L., Kutsyy, V., Coleman, D. A., Cong, G., Crompton, A. M., Elias, K. A., Oestreicher, D. R., Trautman, J. K., and Vaisberg, E. (2006) Compound classification using image-based cellular phenotypes. *Methods Enzymol* **414**, 440–468.
98. Slack, M., Winkler, D., Kramer, J., and Hesterkamp, T. (2009) A multiplexed approach to hit finding. *Curr Opin Drug Discov Devel* **12**, 351–357.
99. Godden, J. W., Furr, J. R., and Bajorath, J. (2003) Recursive median partitioning for virtual screening of large databases. *J Chem Inf Comput Sci* **43**, 182–188.
100. Gough, A. H., and Johnston, P. A. (2007) Requirements, features, and performance of high content screening platforms. *Methods Mol Biol* **356**, 41–61.
101. Niederlein, A., Meyenhofer, F., White, D., and Bickle, M. (2009) Image analysis in high-content screening. *Comb Chem High Throughput Screen* **12**, 899–907.
102. Huang, K., Velliste, M., and Murphy, R. F. (2003) Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. *Proc SPIE* **4962**, 307–318.
103. Young, D. W., Bender, A., Hoyt, J., McWhinnie, E., Chirn, G. W., Tao, C. Y., Tallarico, J. A., Labow, M., Jenkins, J. L., Mitchison, T. J., and Feng, Y. (2008) Integrating high-content screening and ligand-target prediction to identify mechanism of action. *Nat Chem Biol* **4**, 59–68.
104. Tanaka, M., Bateman, R., Rauh, D., Vaisberg, E., Ramachandani, S., Zhang, C., Hansen, K. C., Burlingame, A. L., Trautman, J. K., Shokat, K. M., and Adams, C. L. (2005) An unbiased cell morphology-based screen for new, biologically active small molecules. *PLoS Biol* **3**, e128.
105. Hu, Y., Lounkine, E., and Bajorath, J. (2009) Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function. *ChemMedChem* **4**, 540–548.
106. Nisius, B., Vogt, M., and Bajorath, J. (2009) Development of a fingerprint reduction approach for Bayesian similarity searching based on Kullback-Leibler divergence analysis. *J Chem Inf Model* **49**, 1347–1358.
107. Wang, Y., and Bajorath, J. (2008) Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. *J Chem Inf Model* **48**, 1754–1759.
108. Williams, C. (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. *Mol Divers* **10**, 311–332.
109. Durr, O., Duval, F., Nichols, A., Lang, P., Brodte, A., Heyse, S., and Besson, D. (2007) Robust hit identification by quality assurance and multivariate data analysis of a high-content, cell-based assay. *J Biomol Screen* **12**, 1042–1049.
110. Loo, L. H., Wu, L. F., and Altschuler, S. J. (2007) Image-based multivariate profiling of drug responses from single cells. *Nat Methods* **4**, 445–453.
111. Low, J., Huang, S., Blosser, W., Dowless, M., Burch, J., Neubauer, B., and Stancato, L. (2008) High-content imaging characterization of cell cycle therapeutics through in vitro and in vivo subpopulation analysis. *Mol Cancer Ther* **7**, 2455–2463.
112. Judson, R., Elloumi, F., Setzer, R. W., Li, Z., and Shah, I. (2008) A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinformatics* **9**, 241.
113. Tao, C. Y., Hoyt, J., and Feng, Y. (2007) A support vector machine classifier for recognizing mitotic subphases using high-content screening data. *J Biomol Screen* **12**, 490–496.
114. Wassermann, A. M., Geppert, H., and Bajorath, J. (2009) Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J Chem Inf Model* **49**, 582–592.
115. Kümmel, A., Gubler, H., Gehin, P., Beibel, M., Gabriel, D., and Parker, C. N. (2009) Integration of multiple readouts into the

Z′ factor for assay quality assessment. *J Biomol Screen* **15**, 95–101.
116. MacDonald, M. L., Lamerdin, J., Owens, S., Keon, B. H., Bilter, G. K., Shang, Z., Huang, Z., Yu, H., Dias, J., Minami, T., Michnick, S. W., and Westwick, J. K. (2006) Identifying off-target effects and hidden phenotypes of drugs in human cells. *Nat Chem Biol* **2**, 329–337.
117. Roy, K., and Paul, S. (2010) Docking and 3D-QSAR studies of acetohydroxy acid synthase inhibitor sulfonylurea derivatives. *J Mol Model* **16**, 951–964.
118. Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., Spidlen, J., Strain, E., and Gentleman, R. (2009) flowCore: a bioconductor package for high throughput flow cytometry. *BMC Bioinformatics* **10**, 106.
119. Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L., and Mesirov, J. P. (2009) Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci USA* **106**, 8519–8524.
120. Wermuth, C. G. (2006) Selective optimization of side activities: the SOSA approach. *Drug Discov Today* **11**, 160–164.

Chapter 18

Computational Systems Chemical Biology

Tudor I. Oprea, Elebeoba E. May, Andrei Leitão, and Alexander Tropsha

Abstract

There is a critical need for improving the level of chemistry awareness in systems biology. The data and information related to modulation of genes and proteins by small molecules continue to accumulate at the same time as simulation tools in systems biology and whole body physiologically based pharmacokinetics (PBPK) continue to evolve. We called this emerging area at the interface between chemical biology and systems biology *systems chemical biology* (*SCB*) (Nat Chem Biol 3: 447–450, 2007).

The *overarching goal of computational SCB* is to develop tools for integrated chemical–biological data acquisition, filtering and processing, by taking into account relevant information related to interactions between proteins and small molecules, possible metabolic transformations of small molecules, as well as associated information related to genes, networks, small molecules, and, where applicable, mutants and variants of those proteins. There is yet an unmet need to develop an integrated in silico pharmacology/systems biology continuum that embeds drug–target–clinical outcome (DTCO) triplets, a capability that is vital to the future of chemical biology, pharmacology, and systems biology. Through the development of the SCB approach, scientists will be able to start addressing, in an integrated simulation environment, questions that make the best use of our ever-growing chemical and biological data repositories at the system-wide level. This chapter reviews some of the major research concepts and describes key components that constitute the emerging area of computational systems chemical biology.

Key words: Drug–target–clinical outcome triplets, Pharmacodynamics/pharmacokinetics, Biological networks, Cheminformatics, QSAR modeling, Biochemical network simulations, Systems biology

1. Introduction

Regarded as a departure from the "reductionist approach," where investigators dedicate their efforts to the study of a single gene/protein, *systems biology* (SB) is considered a "comprehensive approach." In SB, large networks describing the regulation of entire genomes, metabolic/transporter, or signal transduction pathways are analyzed in their totality at different levels of biological organization [1]. SB blends theory, computational modeling, and high-throughput experimentation [2], and has

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_18, © Springer Science+Business Media, LLC 2011

already led to advances in cell signaling [3] developmental biology [4], cell physiology [5] and to the understanding of metabolic networks [6]. Recently, we coined the term *systems chemical biology*, which integrates bioinformatic and cheminformatic databases and cheminformatic tools with biological network simulations [7]. We argued that chemistry awareness is required in order to achieve a systematic understanding of the way small molecules affect biological systems. This concept had a positive impact in the chemistry community, as reflected by the 14 papers presented at the SCB symposium organized at the American Chemical Society national meeting in Philadelphia[1] one year later.

Other attempts of utilization of SB technologies include in silico polypharmacology [8, 9] and are deployed in industrial drug discovery [10, 11]. Furthermore, the chemical biology agenda, as embodied by the NIH Roadmap Molecular Libraries Initiative (MLI) [12], enables SCB by extending the study of chemical effects on biological targets toward the entire array of macromolecules and macromolecular networks. These can be further mapped using additional genomic and proteomic tools, in order to gain comprehensive insight into, e.g., phenotypic screening. Via the MLI and its successor, the Molecular Libraries Program (MLP), the effects of hundreds of thousands of small molecules are being investigated on biological systems of varied complexity, from individually screened targets to multiplex screens, phenotypic screens, and other cellular and whole organism assays. Indeed, this unprecedented public effort creates new challenges for advancing chemocentric approaches to systems biology, as increasing amounts of disparate data are being deposited in publicly available databases (*see* Table 1). As of November 13, 2009, PubChem [13] features 328,392 MLP-related compounds, of which, 296,070 are Lipinski's "rule of five" (Ro5) [14] compliant and 152,778 are "active," all tested on 869-MLP related (including 515 "confirmatory") assays, from the high-throughput screening centers network.

This plethora of small molecule data, in addition to those present in other annotated chemical libraries (e.g., WOMBAT) (*see* Table 2) has yet to reach the fields of computational biology and systems biology. As cross-system data related to genes, proteins, and their modulation via diverse libraries of small molecules becomes available, an unmet critical need – chemistry cognizance – is required in order to advance the development of a systems biology, which we believe is vital to the understanding of human health. It is indeed surprising that, with the possible exception of in silico pharmacology [8], none of the computational biology

[1]The symposium "Systems chemical biology: Integrating chemistry and biology for network models" was organized at the 236th ACS National Meeting in Philadelphia, August 17–21, 2008; it was sponsored by CINF and co-sponsored by four other ACS divisions (COMP, MEDI, HEALTH, and BIOT).

Table 1
Public resources for *SCB*[a]

Genes
Entrez gene: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
Proteins
SwissProt: http://expasy.org/sprot/
Structures of biological macromolecules
PDB: http://www.rcsb.org/pdb/home/home.do
Structural genomics consortium: http://www.sgc.utoronto.ca/
Pathways
KEGG: http://www.genome.jp/kegg/
MetaCyc: http://metacyc.org/
BioCarta: http://www.biocarta.com/genes/index.asp
Reactome: http://www.reactome.org/
Receptors
GPCRdb: http://www.gpcr.org/7tm/
NHRs: http://www.nursa.org/
Ion channels: http://www.iuphar-db.org/iuphar-ic/index.html
Biochemical pathway reaction kinetics
SABIORK: http://sabio.villa-bosch.de/SABIORK/
BRENDA: http://www.brenda.uni-koeln.de/
Annotated biological models
http://www.ebi.ac.uk/biomodels/
Other MLI initiatives
NIH Roadmap: http://nihroadmap.nih.gov/

[a]Non-exhaustive list

approaches available to date offers any resolution from a cheminformatics perspective. Cheminformatics, an independent research discipline concerned with the application of information retrieval methods to chemical databases that emerged just over a decade ago [15], has become an integral part in the drug discovery decision-making system [16] and is today the main resource for computer-based studies of chemistry-modulated biological systems [17]. In parallel to the evolution of molecular pharmacology into polypharmacology, cheminformatics is increasingly applied to in silico profile small molecule bioactivities for arrays of targets [8, 9, 18] although it has yet to be fully utilized in chemical biology,

Table 2
Sources of bioactivity data for *SCB*[a]

Small molecules
PubChem: http://pubchem.ncbi.nlm.nih.gov/
NCI: http://dtp.nci.nih.gov/docs/dtp_search.html
WOMBAT: http://sunsetmolecular.com/
BINDING DB: http://www.bindingdb.org/bind/index.jsp
Metabolites: http://www.hmdb.ca/
Drugs and clinical candidates
NLM's Dailymed: http://dailymed.nlm.nih.gov/
DrugBank: http://drugbank.ca/
FDA: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/
WHO essential drugs: http://www.who.int/medicines/publications/essentialmedicines/en/
Toxicology data
NIEHS: http://ntp.niehs.nih.gov/ntpweb/
EPA DSS-Tox: http://www.epa.gov/ncct/dsstox/index.html

[a]Non-exhaustive list

an emerging discipline that aims at modulating all proteins via small molecules [19]. Indeed, without chemistry cognizance, one cannot port cheminformatics predictive tools [16], e.g., virtual screening [20] to systems biology.

The increasing availability of data related to genes, proteins, and their modulation by small molecules creates a critical need to develop systems chemical biology. There is an unmet requirement to develop a cheminformatics interface, which we believe is vital to the future of systems biology and that will enable the prediction of the effects of chemical structures in the context of biological systems. Figure 1 illustrates the complexity of this problem and our vision for the contribution of in silico modeling of chemical structures toward modulation of biological pathways.

Computational systems chemical biology aims to create a computational infrastructure and a platform to predict systemic effects (ultimately including clinical outcome) of an organic compound entering the body via any of the standard routes of administration (oral/i.v./i.m. etc.). To achieve this goal, one should seek to build rigorous PD/PK models to predict such observables as tissue partitioning, half-life, distribution and clearance, ligand–target interaction and drug efficacy, while taking into account the relevant metabolites of a chemical. In addition, one should seek to

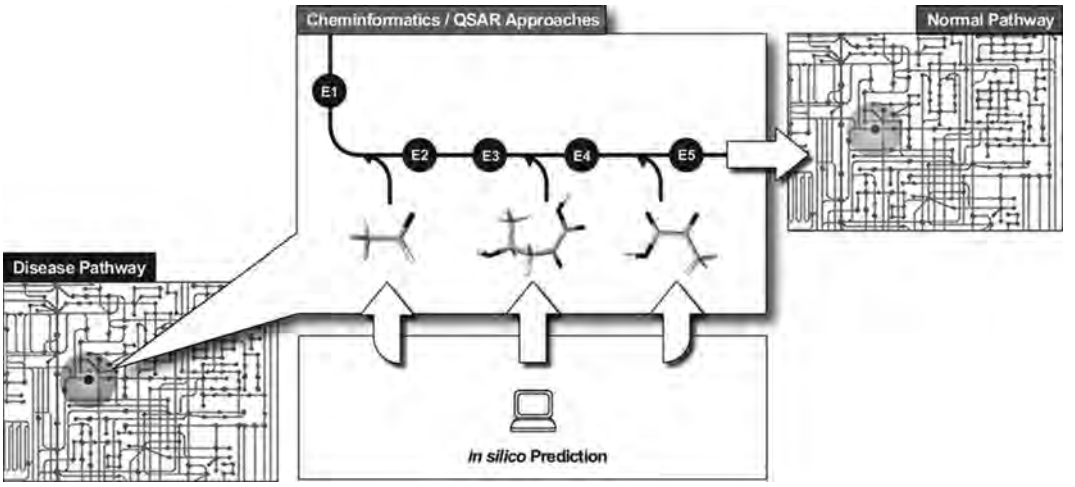


Fig. 1 Contribution of Cheminformatics to Systems Biology. It is expected that computational modeling will afford the prediction of chemical structures active against individual (or multiple) targets while PBPK approaches will afford the estimates of compound distribution and accumulation in target tissues. Yet the knowledge of pathways will enable to predict the effect of chemicals on the entire system in the context of steering the disease-affected network toward a normal state.

predict the specificity of compound interaction with biological targets and simulate the outcome of drug–target interaction at the molecular, cellular, and organ level. The latter objective entails the development of network simulators that explicitly take into account the chemical nature of the small molecules (or their combinations) perturbing the network. This endeavor requires the integration of several complimentary efforts in various fields contributing to the functional SCB workflow incorporating the following tasks: (1) Develop PK/PD models to predict the potential of exogenous small molecules to reach cellular components hosting specific pathways, estimate their concentrations in vivo, and their relationship to specific, understood clinical outcomes; (2) Integrate available data on chemical–target interactions and develop target-specific predictive models of chemical bioactivity using advanced cheminformatics approaches such as Quantitative Structure Activity Modeling (QSAR). These models will enable to predict plausible targets for exogenous compounds from their chemical structure as well as to identify compounds in virtual chemical libraries that are predicted to interact with target proteins and pathways; (3) Investigate, using kinetic network simulation technologies, how small molecules perturb a particular pathway, or perhaps several networked pathways, and predict how these perturbations result in (novel) clinical outcomes. Whereas the comprehensive exploration of SCB requires the consideration of all of the above three major components of the field, we will limit our discussion here to the latter two areas. Several recent reviews provide a lot of detailed information concerning

PK/PD modeling (e.g., [21, 22]); however, in this review we shall consider and illustrate the elements of in silico (multi)target screening and systems biology simulations contributing to the field of SCB.

2. Methods

2.1. SCB Databases: Availability, Compilation, and Curation

Research related to systems biology, chemical probe, and drug discovery produces large amounts of data in seemingly unrelated fields, such as molecular and cellular biology, chemical biology, combinatorial and medicinal chemistry, genetics, and toxicology. This information needs to be organized, queried, and structured to guide the scientific process and to transform data into information and knowledge. Three major components of this process have been identified and discussed elsewhere [23]:

- Chemical and bioactivity information: combines chemical structures with experimental or calculated chemical and physical properties. This type of information relates to the storage of chemical structures and associated molecular data in machine-readable format. Key to storing chemical structures is the atomic connectivity, expressed in connection tables that store two- and/or three-dimensional atomic coordinates. Bioactivity information should capture activity data – primarily activity type and value – with unique indexes identifying the chemical compound, the biological target, cell or organism, with the experimental protocol and bibliographic references. Additional bioactivity fields include experimental observations and errors, images (e.g., Schild plots [24]), as well as keywords such as "partial," "inverse," "competitive," "agonist," "antagonist," and "inhibitor."
- Target and protocol information: biological target and experimental protocol data. This type of information relates to the storage of target and gene information, as well as associated bioassay data in machine-readable format. Many bioinformatics databases are freely available on the Internet. Proper unique identifiers (the equivalent of chemical names), such as those from NCIB/Entrez or Swiss-Prot, enable the end-user to navigate across these databases using uniform resource locators (URL) hyperlinks. Extended target names and functions, as well as information related to their classification and species, will be stored. For example, using functional criteria, a target may be an enzyme (Enzyme Codebook E.C. numbers are stored), a G-protein coupled receptor (GPCR), a nuclear hormone receptor (NHR), an ion-channel, a transporter, or perhaps "other" (unspecified) protein, as well as nucleic acid

(DNA or RNA). The use of a controlled vocabulary should enable data capture and curation of protocol information via predefined keywords, which store information related to specific/nonspecific (radio)ligands, substrates, etc.

- Reference information: bibliographic information for all units in the database. References contain bibliographic information, such as authors or inventors, title, source (e.g., journal name or patent), as well as other pertinent information (volume, page numbers, patent number, etc.). Using unique identifiers, e.g., PubMed or digital object identifiers (DOIs) entries can be hyperlinked to the appropriate abstract or full-text publication via MEDLINE or other databases. Publisher-provided or MeSH (Medline subject headings) keywords can provide further content to the target and protocol fields. In-house reports as well as Internet references should also be indexed, as they provide valuable content.

Computer-based systems for information capture, storage, and retrieval are of critical importance in understanding and mining the systems chemical biology interface. Such information is pertinent to target discovery, to understanding disease models, as well as to the study of bioactive chemotypes, promiscuous scaffolds, and privileged structures. Although the principles for designing the ideal (or desired) SCB databases have been defined as discussed above and primary data is available to a large extent, the comprehensive SCB databases are yet to be established creating a formidable challenge to the field of SCB. The integration process itself requires hierarchical classification schemes since the knowledge related to chemical libraries, biological target families, and biological pathways needs to be mined simultaneously. A variety of chemical, e.g., SciFinder [25] or medicinal chemistry-related databases, e.g., MDDR [26] or drug-related databases such as PDR [26] is available. However, these for-fee databases do not capture critical biological endpoints in numerical form, i.e., there is no searchable field to identify, in a quantitative manner, what is the target- or property-related activity of a particular chemical. This information is important if one considers that (a) not all chemicals indexed in chemical databases are active – some are merely patent claims with no factual basis; and that (b) not all chemicals disclosed as active are equally potent for the target of choice.

To curate SCB data at the appropriate quality level for, e.g., the purpose of understanding pharmacodynamics/pharmacokinetics (PD/PK) models at the molecular levels, it is more appropriate to develop and curate large bioactivity databases. Indeed, all biological research produces large amounts of data that need to be organized, queried and reduced to scientific information and knowledge. Thus, management of biological data involves

acquisition, modeling, storage, integration, analysis, and interpretation of diverse data types. For the purpose of this discussion, biological activity refers to experimentally measured data for a set of chemical compounds on a given biological target (as well as cell, organ, and organism), using predefined experimental protocols. After curation and standardization, these measured values together with related information can be indexed in a bioactivity database. In the largest context, databases need to handle data in a structured and organized way. Consequently, the key task when designing an effective bioactivity database is to properly structure the information. Figure 2 provides an example of the data curation and organization workflow that can be used to design integrated SCB databases.

This model, depicted in Fig. 2, has a two-level structural design (Olah and Oprea 2006). The *internal level* corresponds to the database itself while the *external level* provides cross-referencing support (stored identifiers) for accessing external records from other databases. This database model provides a set of unique and stable identifiers for linking to external levels of other databases. Those databases will perceive this one as external; hence the interconnection through external levels is bidirectional.

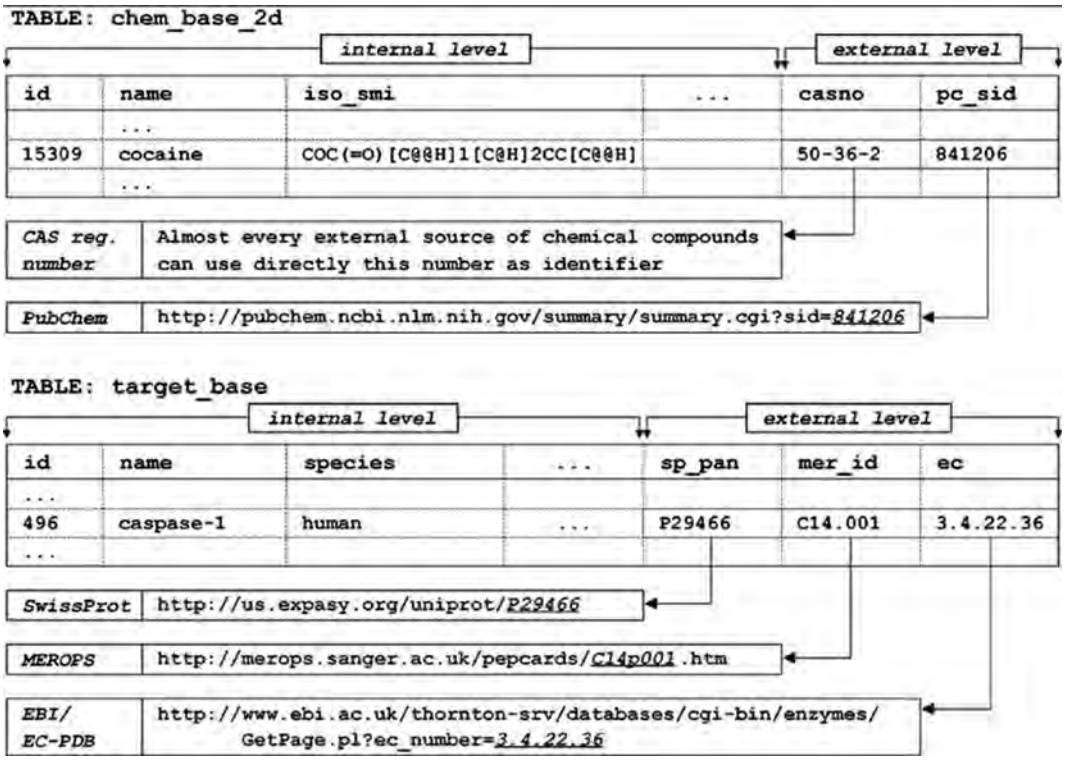


Fig. 2 Data curation workflow.

The creation of specialized SCB databases represents a challenge to be addressed in the near future. Nevertheless, it is of value to discuss at this point several examples of existing databases that would contribute to the desired comprehensive SCB databases.

2.1.1. Complex Bioactivity Databases

To illustrate the complexity and challenges associated with the task of creating chemical biological databases, we could refer to our past experience that includes two databases, namely WOMBAT and WOMBAT-PK [27]. *WOMBAT 2009.1* contains 295,435 entries (242,485 unique SMILES), representing 1,966 unique targets, captured from 14,367 papers published in medicinal chemistry journals between 1975 and 2008. Approximately 61% of these papers are from the American Chemical Society (ACS) journal, J Med Chem; another 30.3% of the papers are from the Elsevier journal, Bioorg Med Chem Lett. Each bioactive molecule has indexed target and bioassay protocol information, with links to the original publication as well as computed chemical descriptors. To date, according to scholar.google.com, WOMBAT has been used as a reference database in over 30 publications related to chemogenomics and medicinal chemistry. *WOMBAT-PK 2009* contains 1,230 entries (1,230 unique SMILES), totaling over 13,000 clinical PK measurements. *WOMBAT-PK 2009* drugs are indexed from multiple literature sources [28, 29], FDA Approved Drug Products [30], peer-reviewed literature, etc.; 1,085 drugs and 36 active metabolites have drug target annotations on 618 targets; an additional 231 drugs are annotated for antitargets [31]. Several physico-chemical property measurements (e.g., water solubility at neutral pH, $LogD_{7.4}$; octanol-water distribution coefficient, LogP; pKa) are also included.

WOMBAT and WOMBAT-PK [27] present examples of databases that we regard as *complex*. Generally speaking, we distinguish two types of complex databases: those that include collections of many cases when a large number of molecules were tested against a single target and those that contain data on a series of compounds tested concurrently in multiple assays. The first type is typically represented by the activity (e.g., Wombat) or "property" datasets (e.g., Wombat-PK, or solubility or toxicity databases) when the property is naturally measured across many molecules. Arguably the largest single collection of such toxicity datasets is *DSSTox* (http://www.epa.gov/nheerl/dsstox/About.html), which includes data such as (a) tumor target site incidence and carcinogenic potencies for 1,354 chemical substances tested in rats and mouse, 80 chemical substances tested in hamsters, 5 chemicals tested in dogs, and 27 chemical substances tested in nonhuman primates; data reviewed and compiled from literature and the National Toxicology Program (NTP) studies; (b) EPAFHM: EPA Fathead Minnow Aquatic Toxicity Database includes Acute

toxicities of 617 chemicals tested in common assay, with mode-of-action assessment and confirmatory measures. In addition, a large collection of single target property datasets is available from http://www.cheminformatics.org/datasets/.

The databases of the second type are rapidly accumulating. The NIH's Molecular Libraries Roadmap Initiative [12] laid out a strategy plan to house information on the biological activities of small molecules (in PubChem [13]) and transform them into chemical probes to perturb specific biological pathways. Currently, PubChem contains more than 25.5 million unique structures for Compound database (of which over 18.3 million are Ro5-compliant) derived from over 60.7 million records in the PubChem Substance database, with links to bioassay description, literature, references, and assay data for each entry. BioAssay Database provides searchable descriptions of nearly 1,918 bioassays, including descriptions of the conditions and readouts specific to a screening protocol. It integrated the vast array of resources, including the 60 Human Tumor Cell lines data from Molecular Targets databases of DTP/NCI and 1,478 MLPCN (Molecular Libraries Probe Production Centers Network) related assays. It is especially useful when chemical information is needed for specific targets, cell lines or diseases. It should be pointed out that the Substance database sourced data information from a multitude of major databases, e.g., Binding Database, ChemBank, NCI/DTP, KEGG, SMID, and ZINC. The Binding Database is a public database of measured binding affinities for biomolecules, containing experimental data of 21,143 binders to 244 biological targets [32]. ChemBank is a suite of informatics tools and databases created by the Broad Institute, aimed at promoting the development of chemical genetics [33]. The Developmental Therapeutics Program (DTP) of the NCI has collected 127,000 compounds in both 2D and 3D formats that are freely available. They were generally screened for evidence of the ability to inhibit the growth of 60 human tumor cell lines over the past 40 years. KEGG (Kyoto Encyclopedia of Genes and Genomes)[2] is an informatics resource for biological systems [34]. It includes four constituent databases, categorized as building blocks in the genomic space (KEGG GENES, 1,720,795 genes), the chemical space (KEGG LIGAND, 14,238 compounds), wiring diagrams of interaction and reaction networks (KEGG PATHWAY, 42,314 pathways), and KEGG BRITE, 5,642 hierarchical classifications. The Small Molecule Interaction Database (SMID) [35] is a database of protein domain-small molecule interactions by using structural

[2]KEGG (http://www.genome.jp/kegg/) has the following entry points: PATHWAY, the KEGG pathway maps for biological processes; BRITE, the KEGG functional hierarchies of biological systems; GENES: the KEGG gene catalogs and ortholog relations in complete genomes; and LIGAND, the KEGG chemical compounds, drugs, glycans, and reactions.

data from the Protein Data Bank (PDB). SMID is essentially a "listing" of all small molecules (5,117 records) that have been shown to bind to any given conserved protein domain (3,508 records), including total 274,917 interactions.

As part of the NIMH Psychoactive Drug Screening Program, PDSP Ki Database (http://pdsp.med.unc.edu/indexR.html) contains 47,458 K_i values, embracing 749 types of receptors and 6,935 test ligands. The majority of the receptors are GPCRs (549 types), along with various enzymes, ion channel and transporters, thus the largest database of its kind in the public domain. As the common observations in GPCRs–ligands interactions, small molecules can bind to multiple set of GPCRs with high affinities. The online data mining tools make it easy to gather the binding profile of ligands and construct the two-dimensional matrix of GPCRs and ligands. An interactive search in iPHACE (Integrative Navigation in Pharmacological Space; http://cgl.imim.es/iphace/), an interactive query system that combines PDSP with the IUPHAR database (http://www.iuphar-db.org/index.jsp), retrieves 25 activities for Ketanserin, a strong binder of 5HT2A receptor: K_i values are available for 11 other 5HT receptors, 5 alpha-adrenoceptors, 4 dopamine receptors, the histamine H1 receptor, the dopamine active transporter, and the serotonin-gated ion channel [36].

In order to be capable of building mathematical models for this complex interaction matrix of multiple targets and ligands, a sophisticated algorithm like multiple objective optimization is indispensable. In summary, this large data warehouse makes possible the mapping of the multidimensional space of GPCRs receptorome and will potentially assist the rational design of these "magic shotgun" ligands. Another GPCR–Ligand Database (GLIDA) is a unique database tailored for GPCR-related chemical genomic research [37]. To date, 3,738 entries of GPCRs are searchable together with 649 ligand entries and 1,989 GPCR–ligand pair entries.

Finally, there are interesting examples of chemogenomics databases that capture the effects of chemicals on gene expression. CEBS Microarray Database, available from the National Center for Toxicogenomics at NIEHS (http://www.niehs.nih.gov/cebs-df/incebs.cfm), provides an integrated solution for searching, analyzing, and interpreting data from several microarray platforms. This is the largest publicly available collection of toxicogenomic data for diverse chemicals including data on toxicogenomic profiles for over 100 chemicals provided by Johnson & Johnson.

2.1.2. Pathway-Specific Databases

Biologically relevant pathways are increasingly available via initiatives such as KEGG, which provides a "complete computer representation of the cell, the organism and the biosphere which will enable computational prediction of higher level complexity of cellular processes and organism behaviors from genomic and

molecular information" [38]. KEGG and other online systems, e.g., BioCyc,[3] BioCarta[4] and Reactome (http://www.reactome.org), summarize vast arrays of data, integrating metabolic, transporter, and signal transduction pathways across a variety of organisms, including humans. These clickable objects lead to additional information related to reactions and pathways, to gene and protein data, e.g., Protein Data Bank [39] for proteins; or to structure names, chemical structures, and other online chemical information, e.g., PubChem [13] and ChemSpider [40] for small molecules, respectively. These network representations of static objects lack dynamic integration. Such dynamic aspects can be incorporated by including temporal components, e.g., kinetics such as the Michaelis–Menten constant (K_M), dissociation rates (K_D) for substrates, or stoichiometric information. Network simulators, based on ordinary differential equations (ODEs) or stochastic methods, are required to make assumptions regarding enzyme/transporter concentrations and reaction velocity, diffusion rates for the appropriate endogenous ligand, as well as the stoichiometry with respect to various partners involved in any given step of the pathway. Figure 3 illustrates a simplified representation of the glyoxylate pathway extracted from KEGG. The chemical structures, added for clarity, as well as other object information, are one click away. With BioXyce (*see* below), such metabolic pathways can be simulated by ensuring that appropriate stoichiometry and metabolic changes (i.e., mass flux) are accounted for.

2.1.3. Bioavailability Databases

The work of Amidon and colleagues [41] was incorporated into the FDA guidance for waiver of in vivo bioavailability and bioequivalence testing of immediate-release solid dosage forms for drugs that are Biopharmaceutics Classification System Class 1 (high-solubility, high-permeability) when such drug products also exhibit rapid dissolution. This guidance reflects the interest of the FDA in decreasing the regulatory burden utilizing a science-led approach. In 2005, Wu and Benet [42] proposed that a Biopharmaceutics Drug Disposition Classification System (BDDCS) could provide a very simple surrogate for permeability: BDDCS Classes 1 and 3 are highly soluble whereas Classes 2 and 4 are poorly soluble; Classes 1 and 2 are extensively metabolized whereas Classes 3 and 4 are poorly metabolized. Wu and Benet suggested that if the major route of elimination for a drug was

[3]BioCyc includes MetaCyc, a database of nonredundant, experimentally elucidated metabolic pathways, that can be queried by Pathway, Reaction and Compound, http://metacyc.org/, and the Open Chemical Database, a collection of associated metabolites, http://biocyc.org/open-compounds.shtml.

[4]Biocarta is a commercially-sponsored "open source" forum that integrates emerging proteomic information from the scientific community and depicts inter-molecular interactions via dynamic graphical models. http://www.biocarta.com/genes/index.asp.

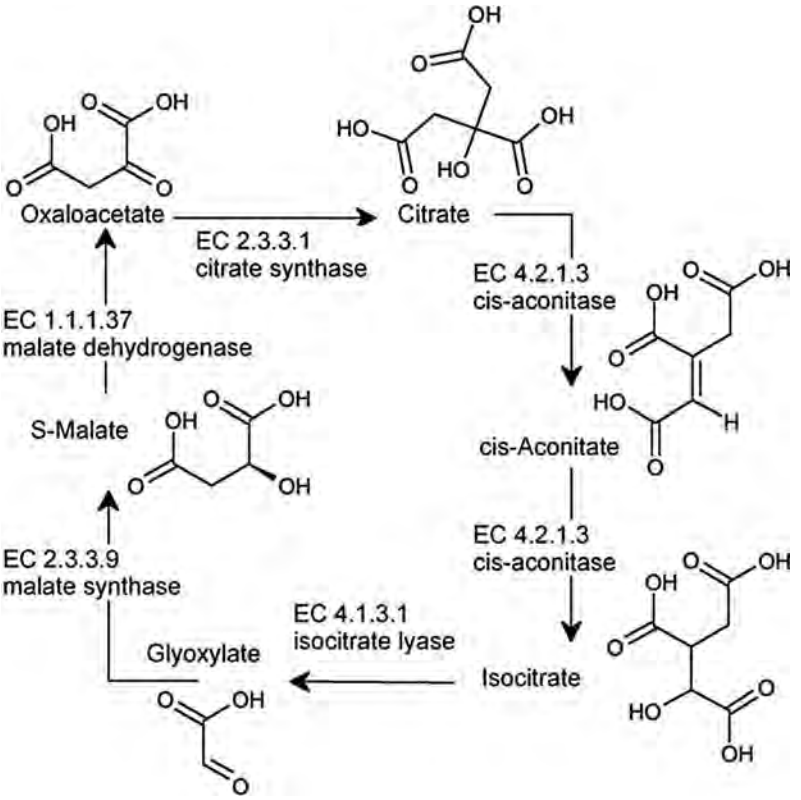


Fig. 3 Glyoxylate pathway (schematic).

metabolism, then the drug exhibited high permeability, while if the major route of elimination was renal and biliary excretion of unchanged drug, then that drug should be classified as low permeability. They further proposed that BDDCS may result in a classification system that yields predictability of in vivo disposition for all four classes, as well as increasing the number of Class 1 drugs eligible for bioequivalence study waivers. This was followed by a recent recommendation [43] that regulatory agencies add the extent of drug metabolism (i.e., ≥90% metabolized) as a method to identify Class 1 drugs suitable for a waiver of in vivo studies of bioequivalence: Following a single oral dose to humans, administered at the highest dose strength, mass balance of the Phase 1 oxidative and Phase 2 conjugative drug metabolites in the urine and feces, measured either as unlabeled, radioactive labeled or nonradioactive labeled substances, accounts for ≥90% of the drug dosed. This is the strictest definition for a waiver based on metabolism. For an orally administered drug to be ≥90% metabolized by Phase 1 oxidative and Phase 2 conjugative processes, it is obvious that the drug must be absorbed [43]. Even 70% metabolism may be appropriate, as suggested in earlier work [42].

Benet and Oprea curated metabolism and solubility information (required for BDDCS classification) for 818 approved drugs and 24 active metabolites, for which human data was available (manuscript in preparation). As metabolites can be excreted in the bile, it is not possible to only use urinary excretion values to validate the extent of metabolism. Given the values for percent excreted unchanged (%Urine) obtained from our curated dataset, many BDDCS Classes 1 and 2 drugs are shown to be ≥70% metabolized: For 277 Class 1 drugs, the median %Urine ± S.D. was 2.0 ± 9.8%, and for 197 Class 2 drugs the values were 1.0 ± 8.8%. By contrast, for 219 Class 3 drugs, the median % Urine ± S.D. was 65 ± 23.6%, and for 39 Class 4 drugs the values were 50 ± 27.1%. Simple cheminformatic analyses based on ClogP (the calculated octanol water partition coefficient) and PSA (the polar surface area) indicate that it is possible to separate BDDCS Class 2 and Class 3 drugs using ClogP and BDDCS Classes 1 and 3 using PSA. These results indicate that filtering tools based on descriptors computed from chemical structures (such as ClogP and PSA) may be used as probability schemes during PK/PD simulations, in particular for Classes 2 and 3, respectively. Although Class 1 drugs do not appear to be influenced by these properties, it should be recalled that efflux transporters do not play a significant effect for these drugs. Furthermore, it is anticipated that building successful in silico models for BDDCS Classes 2 and 3 will assist in giving higher (Class 2) or lower (Class 3) priority for virtual screening for transporters.

2.1.4. Databases Linking Drugs, Targets, and Clinical Outcomes

Current small molecule drugs appear to interact with a rather small number of molecular drug targets: The earlier estimate of 483 targets [44] was recently revised to 218 [45] and 324 [46] targets, respectively. The fact that the number of therapeutic targets is under 500 is surprising considering the size of the "druggable genome" [47], or indeed the size of the human genome itself. More optimistic estimates can be found in, e.g., DrugBank [48], an online resource that indexes 1,678 (of which 1,486 human) targets for 1,485 approved drugs. By definition [45], a drug target is a macro-molecular structure (as defined by molecular mass) that undergoes a specific interaction with therapeutics (i.e., chemicals that are administered to treat or diagnose a disease); the target–drug interaction then results in clinical effect(s). This definition is not always amenable to precision, as exemplified by the following: Hydroxyapatite, a mineral targeted by bisphosphonate drugs such as etidronic acid; Fe and Al, two metals targeted by chelating agents such as deferoxamine; and ammonia, for which intravenous infusion of the amino-acid, arginine, can be used as detoxifying agent.

Earlier attempts at databases linking drugs, targets, and clinical outcomes (DTCO) informatics placed emphasis on the intended drug targets [45], i.e., those targets claimed as being associated with relevant clinical effects by their respective discovery teams, or in the approved drug labels. These targets were considered as "validated" if clinical outcomes correlated in knock-out models, or in vivo observations correlated with in vitro results, e.g., receptor (ant)agonism or enzyme inhibition assays. This minimalistic approach is valid when considering each drug in the context of the intended therapy area. Another study [46] focused on FDA-approved drugs and their targets by including "non-intended" drug targets for, e.g., ritonavir, an HIV-protease inhibitor given in combination with other such inhibitors like lopinavir because it slows down their metabolism via cytochrome P450 3A4 (CYP3A4); ritonavir slows lopinavir breakdown via CYP3A4 inhibition. Thus, CYP3A4 is de facto an intended drug target for ritonavir in the formulation by Abbott, Kaletra™.

Drug side effects extracted from public sources and processed via the COSTART (Concepts of the Coding Symbols for Thesaurus of Adverse Reaction Terms) ontology were recently used to evaluate the probability of two drugs sharing the same drug target given their side effects similarity, for a dataset of 502 drugs and 4,857 known drug–target relations [49]. Of the 13 unexpected drug–target pairs described here, 11 were found to bind to class A aminergic GPCRs and one to the serotonin re-uptake pump (5HTT). By examining a dataset (CEREP Bioprint™) of 2,211 drugs experimentally tested on 188 targets from the same experimental source (CEREP), those five class A GPCR amines and 5HTT, i.e., the targets disclosed for 12 of these 13 "unexpected" findings [49], were found to bind, on average, to over 440 (out of 2,211) small molecules. This renders these drug targets "promiscuous" (i.e., ~20% probability of binding to small molecule drugs). Furthermore, we were unable to confirm 5 of these 13 activities in the same CEREP dataset (2007 release; the access to CEREP was kindly provided to one of the co-authors (TIO) by Dr. Scott Boyer from AstraZeneca). While we do not question the methodology of this paper [49], we illustrate that such discrepancies make it difficult to collect reliable information (e.g., CEREP Bioprint™ may have incorrect data). Key to the DTCO triplet annotation is our own prior work, i.e., the annotation of small molecules to targets as indexed in WOMBAT and WOMBAT-PK, two manually curated databases [27]. For example, WOMBAT-PK annotates 1,136 drugs on 618 unique drug targets and antitargets. These elements are prerequisite for successful DTCO triplet identification. Using WOMBAT-PK, we found 3,053 potential DTCO triplets; however, these are not unique. Furthermore, "antitarget" refers to a drug target that is associated with a

significant side effect (e.g., anticancer drugs are substrates for the ATP-binding cassette transporters, such as ABCB1, which cause multidrug resistance in tumor cells).

2.2. Computational Approaches for Modeling SCB Data to Predict Drug–Target Associations

As discussed in the introductory part of this chapter, the SCB investigation of a compound entails answering three major questions: whether it would interact with specific target(s); whether it will reach the target(s); and what pathway (network) will be perturbed when a compound will interact with its targets. Cheminformatics approaches are most useful in addressing the first issue; thus, models that link structure and activity of molecules against specific targets using historic data can be used prospectively to make plausible assertions about specific target activity for new molecules.

There are several computational approaches that can be employed to predict novel compound–target associations. Structure based virtual screening has become a fundamental part of modern computer aided drug design [50, 51]. It entails docking and scoring libraries of small molecules to find compounds that fit into the binding site and bind tightly to the receptor. Since the first seminal publication by the Kuntz group in 1982 [52], this approach has been used successfully in numerous studies resulting in many cases (such as HIV protease inhibitors) in the design of approved drugs [53]. Numerous algorithms and programs have been introduced (for reviews *see* [54–56]). The examples of widely used docking programs include Dock [57], FlexE, and Gold [58].

Traditional docking protocols and scoring functions rely on explicitly defined three dimensional coordinates and standard definitions of atom types of both receptors and ligands. Albeit reasonably accurate in some cases, structure-based virtual screening approaches are for the most part computationally inefficient, which limits the size of computationally tractable screening compound collections. Furthermore, recent extensive studies into the comparative accuracy of multiple available scoring functions suggest that accurate prediction of binding orientations and affinities of receptor ligands remains a formidable challenge (e.g., [59]). Finally, the number of targets with well-characterized crystal structure that could be used for virtual screening is relatively small compared to the number of targets and assays that have been annotated in ligand databases discussed in Subsection 1 and Table 2. Structure based approaches could and should be considered as a means of predicting chemical–target associations in the context of SCB when feasible. However, since this book focuses on cheminformatics methodologies in general, we will not discuss the structure-based virtual screening methods in detail here; good description of these approaches could be found in the literature includes several publications cited above.

Cheminformatics approaches based on concepts of chemical similarity, pharmacophore or QSAR modeling are finding growing applications as virtual screening tools. Many of these approaches have been reviewed in a recent specialized monograph [20]. Typical methods rely on representing compounds by multiple chemical descriptors and using chemical similarity algorithms of varying complexity to assert the association between a molecule and a target based on the argument that this molecule is similar to known ligands of the target. Perhaps one of the most interesting approaches in this category was developed recently in B. Shoichet group at UCSF. The method called the Similarity Ensemble Approach (SEA) [60] is based on the estimation of the relative similarity between a new compound and precomputed clusters of molecules with known pharmacology. The association with a target is predicted based on the pharmacological annotation of a cluster with the highest similarity to a query molecule. This approach was recent reported to lead to several significant experimentally confirmed predictions [61].

The similarity or pharmacophore-based approaches predict target–ligand association at the qualitative level. However, in SCB application, it is desirable to predict the ligand–target binding affinity quantitatively because the predicted binding affinity value could be used in SCB network simulations discussed below. Such predictions can be afforded by QSAR models that we shall consider in detail here.

Modern QSAR modeling is a very complex and complicated field requiring deep understanding and thorough practicing to develop robust models. Multiple types of chemical descriptors and numerous statistical model development approaches can be found in specialized literature and so need not be discussed in this chapter. Instead, we shall present several unifying concepts that underlie practically any QSAR methodology especially in the context of prospective use of models for virtual screening. Modern QSAR approaches are characterized by the use of multiple descriptors of chemical structure combined with the application of both linear and nonlinear optimization approaches, and a strong emphasis on rigorous model validation to afford robust and predictive models. The most important recent developments in the field concur with a substantial increase in the size of experimental datasets available for the analysis and an increased application of QSAR models as virtual screening tools to discover biologically active molecules in chemical databases and/or virtual chemical libraries [62, 63]. The latter focus differs substantially from the traditional emphasis on developing so-called explanatory QSAR models characterized by high statistical significance but only as applied to training sets of molecules with known chemical structure and biological activity.

Our experience suggests that QSAR is a highly experimental area of statistical data modeling where it is impossible to decide a priori as to which particular QSAR modeling method will prove most successful. To achieve QSAR models of the highest internal, and most importantly, *external* accuracy, the combi-QSAR approach explores all possible binary combinations of various descriptor types and optimization methods along with external model validation. Each combination of descriptor sets and optimization techniques is likely to capture certain unique aspects of the structure–activity relationship. Since our ultimate goal is to use the resulting models as reliable activity (property) predictors, application of different combinations of modeling techniques and descriptor sets will increase our chances for success.

In our critical publications [64, 65], we have recommended a set of statistical criteria, which must be satisfied by a predictive model. For continuous QSAR, criteria that we will follow in developing activity/property predictors are as follows: (1) correlation coefficient R between the predicted and observed activities; (2) coefficients of determination [66] (predicted vs. observed activities R_0^2, and observed vs. predicted activities R'^2_0 for regressions through the origin); (3) slopes k and k' of regression lines through the origin. We consider a QSAR model *predictive*, if the following conditions are satisfied:

1. $q^2 > 0.5$
2. $R^2 > 0.6$
3. $\frac{(R^2 - R_0^2)}{R^2} < 0.1$ and $0.85 \leq k \leq 1.15$ or $\frac{(R^2 - R'^2_0)}{R^2} < 0.1$ and $0.85 \leq k' \leq 1.15$
4. $\left| R_0^2 - R'^2_0 \right| < 0.3$

where q^2 is the cross-validated correlation coefficient calculated for the training set, but all other criteria are calculated for the test set.

Figure 4 summarizes our overall QSAR modeling strategy that is focused on delivering validated predictive models and ultimately, identification of computational hits predicted to interact with specific targets. We start by randomly selecting a fraction of compounds (typically, 10–15%) as an external evaluation set. The remaining compounds are then divided rationally (using the Sphere Exclusion protocol developed in our laboratory [67]) into multiple training and test sets that are used for model development and validation, respectively using criteria discussed in more detail below. We employ multiple QSAR techniques based on combinatorial exploration of all possible pairs of descriptor sets coupled with various statistical data mining techniques and select models characterized by high accuracy in predicting both training

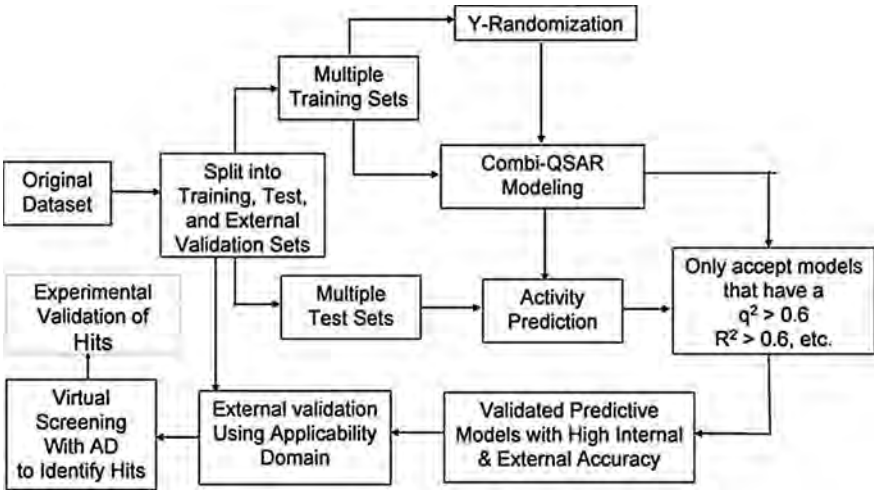


Fig. 4 Flowchart of predictive QSAR modeling workflow implementing combinatorial QSAR modeling and extensive model validation procedures.

and test sets data. Validated models are finally tested using the evaluation set. The critical step of the external validation is the use of applicability domains. If external validation demonstrates significant predictive power of the models, we use all such models for virtual screening of available chemical databases (e.g., ZINC [68]) to identify putative active compounds and seek collaborators who could validate such hits experimentally. The entire approach is described in detail in several recent papers and reviews (e.g., [63]).

We shall note that our approach shifts the emphasis on ensuring good (best) statistics for the model that fits known experimental data toward generating testable hypothesis about purported bioactive compounds. *Thus, the output of the modeling has exactly same format as the input, i.e., chemical structures and (predicted) activities making model interpretation and utilization completely seamless for medicinal chemists.* Note that since we cannot generally guarantee that every prediction resulting from our modeling effort will be validated experimentally, we cannot include the experimental validation step as a mandatory part of the workflow on Fig. 2, which is why we used the dotted line for this component. Nevertheless, in several recent collaborative studies, we have reported on the discovery of experimentally confirmed compounds active against a variety of enzymes and receptors (e.g., [69–73]). These recent successes indicate the power of the predictive QSAR modeling workflow (Fig. 4) as a reliable tool for accurate quantitative prediction of novel ligand–target associations and respective binding constants. Thus, the progressive modeling of all available target bioactivity databases such as those considered in Subsection 3.1 and summarized in our review

[23], which is ongoing in our laboratory, will result in a library of models covering the currently characterized SCB space. Profiling any new compound against this library would result in assigning this compound to one (or may be few) of the target classes (provided that the compound is within the applicability domains of respective target-specific models) and predicting its binding affinity that can be used as a parameter in network simulation models considered in the next section.

2.3. Biological Network Simulations

Due to a growing interest of research community to system-wide understanding and simulations of biological effects, several approaches have been reported [74–80]. To illustrate the capability of a network simulator, we shall briefly describe BioXyce [80, 81], a biological network modeling tool based on Xyce, a massively parallel electrical circuit modeling tool developed by Sandia National Laboratories/DOE (Department of Energy). At the cellular level, biological networks are modeled as electrical circuits where signals are produced, propagated, and sensed. BioXyce uses the following equivalents: chemical mass as charge, mass flux as electric current, concentration as voltage, stoichiometric conservation as Kirchhoff's voltage law, and mass conservation as Kirchhoff's current law. With BioXyce, one can simulate large networks consisting of entire cells, homogeneous cell cultures, or heterogeneous interacting host-pathogen systems in order to understand the dynamics and stability of such systems. To address the challenge of ambiguous rate parameters, BioXyce input parameters, collected from literature, are optimizable using empirical data and the DAKOTA (Design Analysis Kit for Optimization and Terascale Applications) UQ (uncertainty quantification) toolkit [82] (http://www.cs.sandia.gov/DAKOTA/index.html). We can further augment the BioXyce/DAKOTA framework using computational reachability techniques to set initial value conditions and provide tighter parameter bounds [83]. This results in bionetwork models able to replicate behaviors consistent with known experimental data, as shown in Fig. 5 [83]. BioXyce can be used to model and simulate relevant metabolic transporter and signal transduction pathways. Such simulations gain in accuracy by incorporating reaction kinetics data such as K_M and K_{cat}, (the turnover rate), both available from the Comprehensive Enzyme Information System BRENDA [84, 85]. As an illustration, let us consider an SCB analysis of a latency-dependent pathway of *Mycobacterium tuberculosis* (Mtb).

2.3.1. Data Collection

M. tuberculosis is able to persist in host tissues in a nonreplicating persistent (NRP), or latent state. This presents a challenge in the treatment of TB (tuberculosis). Latent TB can reactivate in ~10% of individuals with normal immune systems, higher for those with compromised immune systems. To develop an effective treatment

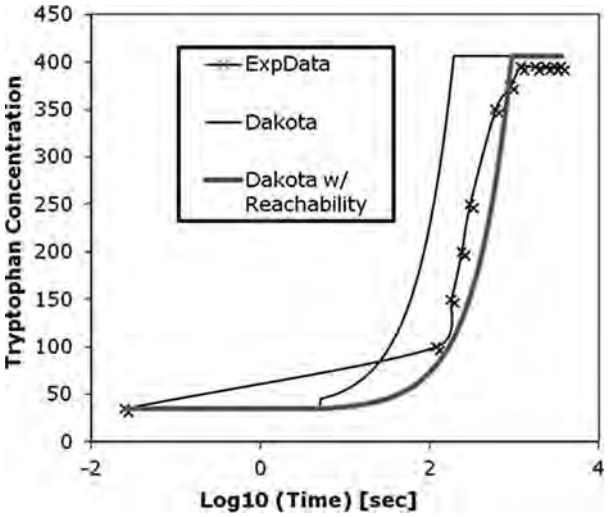


Fig. 5 BioXyce simulation of tryptophan biosynthesis; comparison to experimental data.

against latent TB, we need to understand how potential antimicrobial agents may affect NRP Mtb. We investigated the hypoxic model and virulence associated pathways. In the hypoxic model of NRP, the tubercle bacilli can circumvent the shortage of oxygen by developing alternative energy generation mechanisms. It has been observed that during anaerobic growth conditions, isocitrate lyase (ICL) increases fivefold [86]. ICL is the first enzyme in the glyoxylate bypass pathway (*see* also Fig. 5). Since a comparable increase in the second enzyme in the pathway, malate synthase, is not observed, it was hypothesized that ICL increase with the subsequent glyoxylate increase may serve to replenish NAD by way of the glyoxylate-to-glycine (GtG) shunt [86, 87]. Thus, the interruption of reactions involved in the GtG shunt may prove a means to combat latent Mtb. To demonstrate the feasibility of using SCB to analyze virulence-relevant pathways, we investigate malate synthase inhibition: Although ICL is the critical enzyme, ICL inhibitors are not readily identifiable.

2.3.2. Pathway Simulation

Using data from BioCyc and the averaged reaction rates derived from BRENDA, we simulated the glyoxylate cycle pathway (Fig. 5), involved in the GtG shunt. Reaction rates for each enzyme in the pathway are estimated based on values from the BRENDA database. Stoichiometric relations and enzyme rates were used to construct a biological netlist using BioXyce. BioXyce enables simulations and analyses of whole cell and multicellular systems; this is likely to facilitate the exploration of potential side effects of pathway specific perturbations on nontarget pathways. However, introducing perturbing ligands on any systems biology

network can only be simulated by a break in the circuit that does not take into account any specifics related to the small molecule per se. This lack of chemistry awareness can be addressed by integrating cheminformatics tools, as outlined below.

2.3.3. SCB-Related Virtual Screening Studies

To enable chemistry cognizance in the Mtb pathway simulations, we applied virtual screening to support SCB simulations in the identification of small molecule bioactives – a process that could be used to support PK/PD and FS. We took advantage of the presence of 3D structures (from X-ray crystallography) for two of the enzymes in the Mtb glyoxylate shunt, namely malate synthase and ICL. The substrate binding site in each enzymes was evaluated using GRID [88]: ICL has a very polar binding site that accommodates three carboxylate moieties (data not shown); this makes it unsuitable for small molecule drugs, in particular since drugs require a certain degree of permeability (i.e., non-polar) in order to pass not only the intestinal and/or cellular membranes, but also the Mtb walls as well. Malate synthase (PDB entry 2GQ3) has an active site that accommodates the substrate (glyoxylate) and the cofactor, acetyl-CoA, in order to release malate and CoA (Fig. 6a, b). This site, already subjected to investigation [89], features a relatively small number of hydrophobic interactions (Fig. 6c), which suggests that classical inhibitor design methods may prove unsuccessful. However, we detected another cavity in the vicinity of the catalytic site (Fig. 6d, magenta), which may function as an allosteric site and is more hydrophobic. Preliminary docking studies (with FRED from OpenEye) correctly placed malate in the malate synthase binding site using 2GQ3 and keeping Mg^{2+} and four water molecules. Itaconate, a weak inhibitor, appears to bind like malate and does not dock in the allosteric site. Although we did not find potent ligands targeting the allosteric site, we expect this to become an interesting site for small molecules. No allosteric modulators of malate synthase have been described to date.

2.3.4. SCB Simulation Results

Simulations in the absence of inhibitory molecules were conducted and compared to simulations in the presence of inhibitory molecules identified through the use of cheminformatics analysis tools, as previously described. The current simulation uses a simple competitive inhibition model, where the K_M is increased by $[I]/K_i$ ([I] is the concentration of the inhibitor and K_i is the inhibition constant). The simulation framework allows the incorporation of more complex inhibition models. The incorporation of inhibitors of malate synthase should directly affect the accumulation of glyoxylate and malate. Figure 7 shows these two metabolites in the presence and absence of various inhibitors (Table 3) and differing concentrations of bromopyruvate. Simulation results for the noninhibited system verified that as glyoxylate

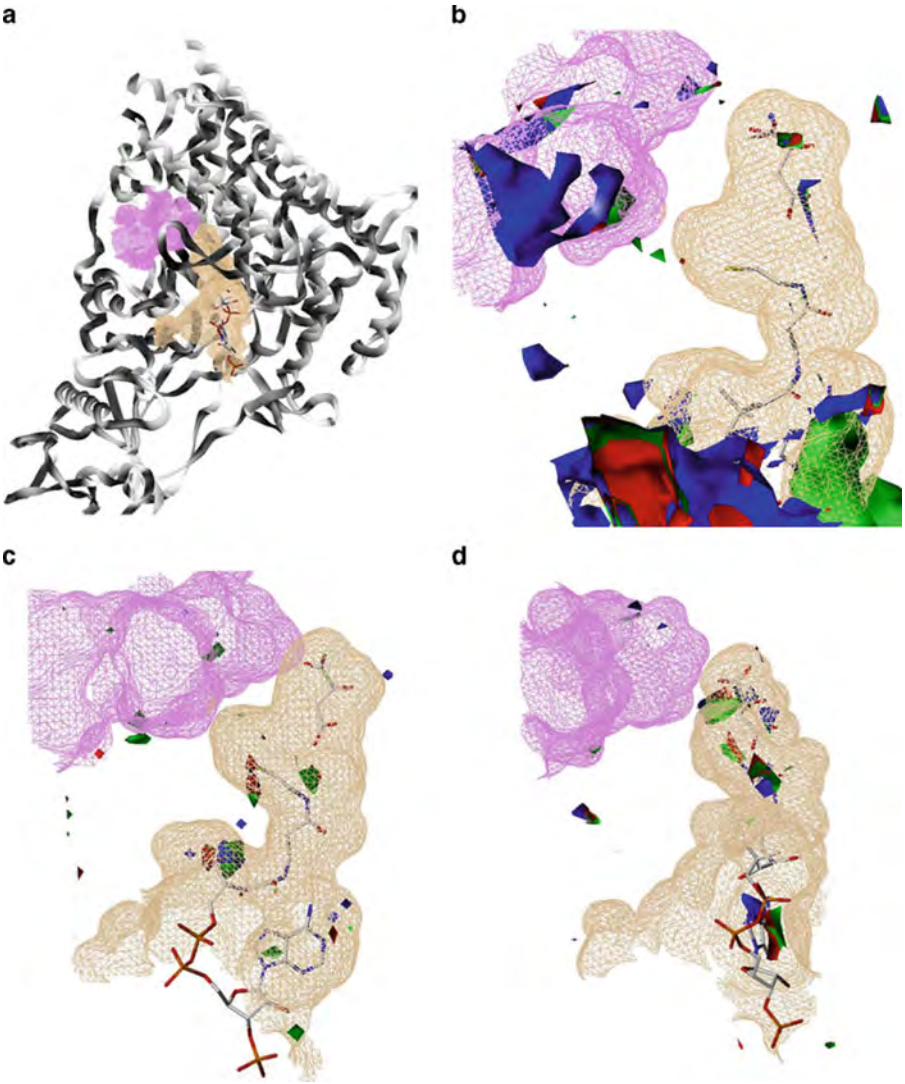


Fig. 6 The malate synthase cavities.

accumulates, malate is produced and eventually consumed to produce downstream metabolites. In the presence of various inhibitory molecules (Table 3), glyoxylate accumulates at a much higher level than the uninhibited state; malate is consumed and not produced at the same rate as in the uninhibited pathway. Combining the simulation platform with the SCB analysis, we observed differences between weak and strong inhibitors and differences in dosage for 1 mM versus 10 mM of Bromopyruvate, thus demonstrating the possibility for an SCB-based approach to probing virulence-relevant pathways.

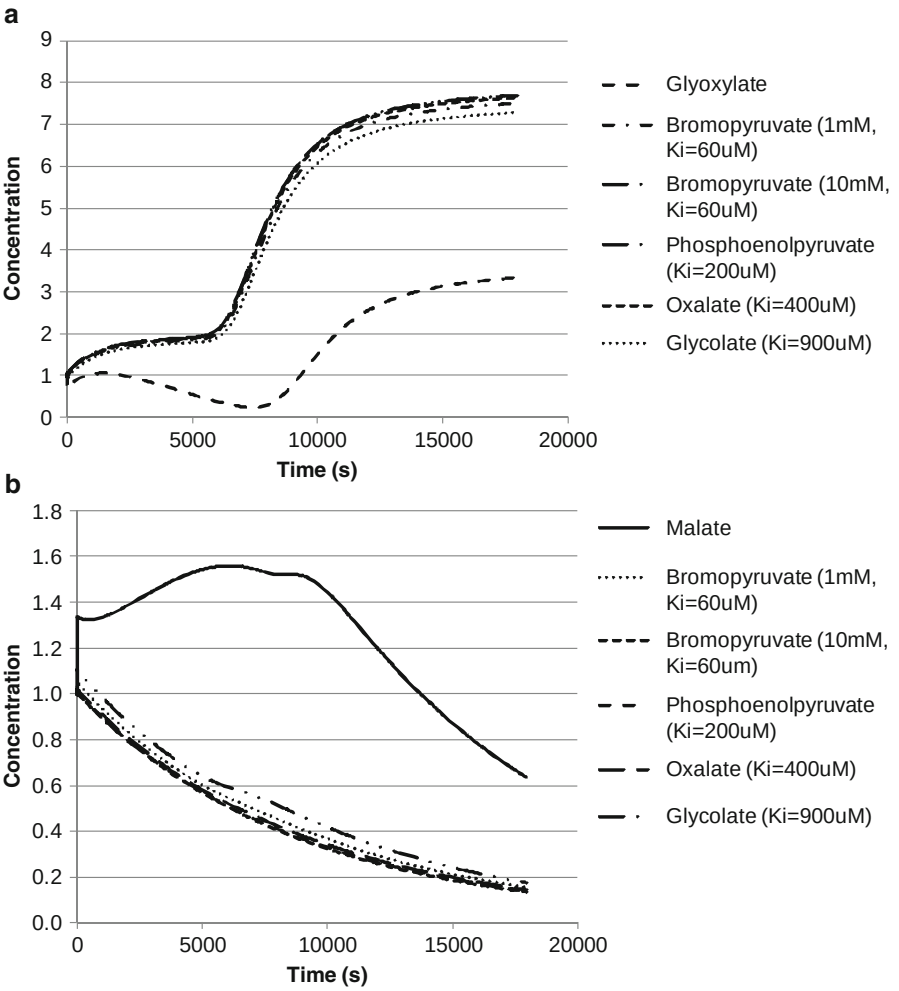
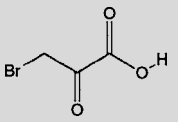
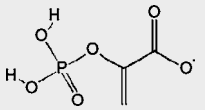
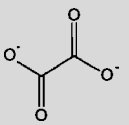
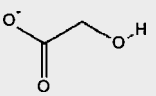


Fig. 7 Glyoxylate (**a**) and malate (**b**) in the presence and absence of inhibitory molecules.

2.3.5. Integration of Network Simulations and Cheminformatics

The future development of SCB will inevitably include the integration of biological network simulations and results of cheminformatics investigation of ligand–target databases. We shall illustrate possible scenarios using studies planned by our group of coauthors around the BioXyce simulator (Fig. 8). Stoichiometric equations can be derived using publicly available databases (Table 1). Reaction rates for each protein in the pathway can be compared, and curated, using BRENDA and SABIORK for enzymes, and other online resources for signal transduction pathways (e.g., from IUPHAR). The stoichiometric relation and enzyme rates can be used to generate biological netlists (native format input for Xyce, the simulation engine of BioXyce). Simulations in the absence of inhibitory molecules can be conducted and verified for internal consistency. A challenge in the development

Table 3
Malate synthase ligands

Compound	Structure	K_i (μM)
Bromopyruvate (inhibitor)		60
Phosphoenol-pyruvate (weak inhibitor)		200
Oxalate (weak inhibitor)		400
Glycolate (very weak inhibitor)		900

of accurate biological network simulations for SCB is the availability of accurate rate data. To this end, we can couple the BioXyce netlist pathways to the DAKOTA optimization environment to find the optimal rate constants that increase the phenotypic accuracy of our simulation. Based on an error analysis, DAKOTA generates new values for the rate constants, which are incorporated into a parameter file included in the BioXyce netlist. Iterative cycles of the BioXyce-DAKOTA coupling help determine unknown rates for pathways of interest.

Whenever perturbing ligand data becomes available, we can compare validated models with simulations in the presence of perturbing molecules. For example, in the case of enzyme inhibitors, we can assume simple competitive inhibition models, where the K_M will increase by $[I]/K_i$ (where [I] is the concentration of the inhibitor and K_i is the inhibition constant). For receptor-based models, we can assume the equivalent of the Michaelis–Menten kinetics and models, such as the Ariens, Simonis, and de Groot [90] models, for intrinsic activity. If needed, we can further take into effect transducer kinetics (such as G-protein coupling),

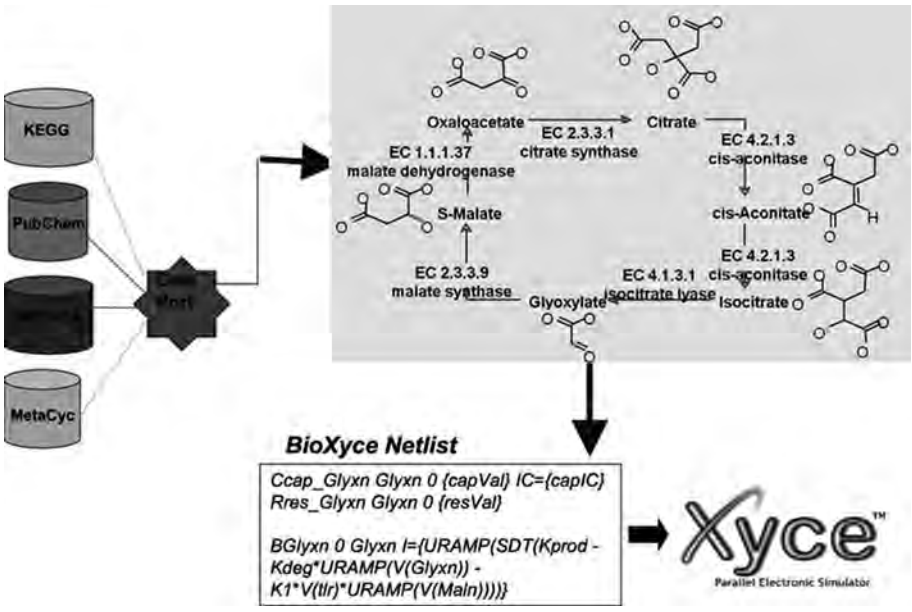


Fig. 8 BioXyce workflow: information from various data sources is integrated and transferred to Xyce input for biological network simulation. The Mtb glyoxylate pathway is depicted.

as well as the observed pharmacological effects (agonism, antagonism, inverse agonism).

In addition, we could think of several extensions of the BioXyce/DAKOTA model. Given empirical data, BioXyce/DAKOTA models are developed in the absence of small molecule perturbants, and the production of metabolites is typically compared to known (observed) outcomes. Models can then be extended to incorporate knowledge related to small molecules and their influence on the model system. We can then simulate the system assuming the presence of ligands that interfere with key enzymes in the pathway. It is anticipated that interference with critical enzyme(s) will reduce the concentration of key metabolites compared to normal. At this stage, the model will have the ability to incorporate output from cheminformatics.

3. Conclusions

The development of integrated systems chemical biology interface could dramatically alter our way of thinking about complex biological networks and unlock the true potential of in silico chemical biology studies of cellular and organism functions. By gaining access to the "known" as well as the "predictive" aspects of small molecule-biological network interactions, scientists could be guided to understand, for example, the potential therapeutic

impact of a small-molecule blockade of a critical step in a pathway. This may ultimately allow an understanding of why some but not all proteins within a pathway make good drug targets, and it may encourage an early focus on those targets that are the most likely to be clinically useful. We anticipate that the emerging field of computational systems chemical biology will see many important advances and discoveries in near future.

Acknowledgments

The authors would like to acknowledge the support of their studies from NIH (grants R01GM066940 and R21GM076059 supporting AT and 1U54MH084690-01 supporting TIO).

References

1. Voit E., Neves A. R., and Santos H. (2006) The intricate side of systems biology. *Proc Natl Acad Sci U S A* **103**, 9452–9457.
2. Kell D. B. (2006) Theodor Bucher Lecture. Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells. Delivered on 3 July 2005 at the 30th FEBS Congress and the 9th IUBMB conference in Budapest. *FEBS J* **273**, 873–894.
3. Blinov M. L., Faeder J. R., Goldstein B., and Hlavacek W. S. (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *Biosystems* **83**, 136–151.
4. Ochi H. and Westerfield M. (2007) Signaling networks that regulate muscle development: lessons from zebrafish. *Dev Growth Differ* **49**. 1–11.
5. Brandman O., Ferrell J. E., Jr., Li R., and Meyer T. (2005) Interlinked fast and slow positive feedback loops drive reliable cell decisions. *Science* **310**, 496–498.
6. Covert M. W., Knight E. M., Reed J. L., Herrgard M. J., and Palsson B. O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96.
7. Oprea T. I., Tropsha A, Faulon J. L., and Rintoul M. D. (2007) Systems chemical biology. *Nat Chem Biol* **3**, 447–450.
8. Mestres J., Martin-Couce L., Gregori-Puigjane E., Cases M., and Boyer S. (2006) Ligand-based approach to in silico pharmacology: nuclear receptor profiling. *J Chem Inf Model* **46**, 2725–2736.
9. Paolini G. V., Shapland R. H., van Hoorn W. P., Mason J. S., and Hopkins A. L. (2006) Global mapping of pharmacological space. *Nat Biotechnol* **24**, 805–815.
10. Morphy R. and Rankovic Z. (2007) Fragments, network biology and designing multiple ligands. *Drug Discov Today* **12**, 156–160.
11. Loging W., Harland L., and Williams-Jones B. (2007) High-throughput electronic biology: mining information for drug discovery. *Nat Rev Drug Discov* **6**, 220–230.
12. Austin C. P., Brady L. S., Insel T. R., and Collins F. S. (2004) NIH Molecular Libraries Initiative. *Science* **306**, 1138–1139.
13. PubChem. (2009) http://pubchem.ncbi.nlm.nih.gov/.
14. Lipinski C. A., Lombardo F., Dominy B. W., and Feeney P. J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3–26.
15. Brown F. (2005) Editorial opinion: chemoinformatics – a ten year update. *Curr Opin Drug Discov Devel* **8**, 298–302.
16. Olsson T. and Oprea T. I. (2001) Cheminformatics: a tool for decision-makers in drug discovery. *Curr Opin Drug Discov Devel* **4**, 308–313.
17. Willett P. (2008) A bibliometric analysis of the literature of chemoinformatics. *Aslib Proc* **60**, 4–17.
18. Fliri A. F., Loging W. T., Thadeio P. F., and Volkmann R. A. (2005) Biological spectra analysis: linking biological activity profiles to

molecular structure. *Proc Natl Acad Sci U S A* **102**, 261–266.
19. Schreiber S. L. (2005) Small molecules: the missing link in the central dogma. *Nat Chem Biol* **1**, 64–66.
20. Varnek A. and Tropsha A. (2008) Cheminformatics Approaches to Virtual Screening. London: RSC.
21. Danhof M., de Lange E. C., Della Pasqua O. E., Ploeger B. A., and Voskuyl R. A. (2008) Mechanism-based pharmacokinetic-pharmacodynamic (PK-PD) modeling in translational drug research. *Trends Pharmacol Sci* **29**, 186–191.
22. Schmidt S., Barbour A., Sahre M., Rand K. H., and Derendorf H. (2008) PK/PD: new insights for antibacterial and antiviral applications. *Curr Opin Pharmacol* **8**, 549–556.
23. Oprea T. and Tropsha A. (2006) Target, chemical and bioactivity databases – integration is key. *Drug Discov Today* **3**, 357–365.
24. de Jong L. A., Uges D. R., Franke J. P., and Bischoff R. (2005) Receptor-ligand binding assays: technologies and applications. *J Chromatogr B Analyt Technol Biomed Life Sci* **829**, 1–25.
25. SciFinder: American Chemical Society, *CAS online/SciFinder*. (2009) http://www.cas.org/SCIFINDER/.
26. MDDR.SYMYX technologies. (2009) http://www.mdl.com/products/knowledge/drug_data_report/index.jsp.
27. Olah M., Rad R., Ostopovici L., Bora A., Hadaruga N., Hadaruga D. et al. (2007) WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery. In: Schreiber S. L., Kapoor T. M., Weiss G. (Eds). Chemical Biology: From Small Molecules to Systems Biology and Drug Design. New York: Wiley-VCH, 760–786.
28. Brunton L, Lazo J, Parker K.(Eds). (2006) Goodman and Gilman's The Pharmacological Basis of Therapeutics. 11th ed. New York: McGraw-Hill; 1984 pp.
29. Physicians' Desk Reference (2009) 63rd ed. PDR., Montvale, NJ, Thomson Reuters, 3315 pp.
30. FDA Approved Drug Products (2009) Available from http://www.accessdata.fda.gov/scripts/cder/ob/default.cfm
31. Vaz R. and Klabunde T. (2008) Antitargets: prediction and prevention of drug side effects. In: Methods and Principles in Medicinal Chemistry. Weinheim: Wiley-VCH.
32. Chen X., Liu M., and Gilson M. K. (2001) BindingDB: a web-accessible molecular recognition database. *Comb Chem High Throughput Screen* **4**, 719–725.
33. Strausberg R. L. and Schreiber S. L. (2003) From knowing to controlling: a path from genomics to drugs using small molecule probes. *Science* **300**, 294–295.
34. Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K. F., Itoh M., Kawashima S. et al. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**(Database issue), D354–D357.
35. Snyder K. A., Feldman H. J., Dumontier M., Salama J. J., and Hogue C. W. (2006) Domain-based small molecule binding site annotation. *BMC Bioinformatics* **7**, 152.
36. Garcia-Serna R., Ursu O., Oprea T., and Mestres J. (2010) iPHACE: integrative navigation in pharmacological space. *Bioinformatics.* **26**, 985–986.
37. Okuno Y., Yang J., Taneishi K., Yabuuchi H., and Tsujimoto G. (2006) GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res* **34** (Database issue), D673–D677.
38. KEGG. (2009.) http://www.genome.jp/kegg/.
39. Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H. et al. (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
40. ChemSpider. (2009) ChemSpider. http://www.chemspider.com.
41. Amidon G. L., Lennernas H., Shah V. P., and Crison J. R. (1995) A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharm Res* **12**, 413–420.
42. Wu C. Y. and Benet L. Z. (2005) Predicting drug disposition via application of BCS: transport/absorption/elimination interplay and development of a biopharmaceutics drug disposition classification system. *Pharm Res* **22**, 11–23.
43. Benet L. Z., Amidon G. L., Barends D. M., Lennernas H., Polli J. E., Shah V. P. et al. (2008) The use of BDDCS in classifying the permeability of marketed drugs. *Pharm Res* **25**, 483–488.
44. Drews J. and Ryser S. (1997) The role of innovation in drug development. *Nat Biotechnol* **15**, 1318–1319.
45. Imming P., Sinning C., and Meyer A. (2006) Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov* **5**, 821–834.
46. Overington J. P., Al Lazikani B., and Hopkins A. L. (2006) How many drug targets are there? *Nat Rev Drug Discov* **5**, 993–996.

47. Hopkins A. L. and Groom C. R. (2002) The druggable genome. *Nat Rev Drug Discov* **1**, 727–730.
48. Wishart D. S., Knox C., Guo A. C., Shrivastava S., Hassanali M., Stothard P. et al. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34** (Database issue), D668–D672.
49. Campillos M., Kuhn M., Gavin A. C., Jensen L. J., and Bork P. (2008) Drug target identification using side-effect similarity. *Science* **321**, 263–266.
50. Brooijmans N. and Kuntz I. D. (2003) Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct* **32**, 335–373.
51. Kitchen D. B., Decornez H., Furr J. R., and Bajorath J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**, 935–949.
52. Kuntz I. D., Blaney J. M., Oatley S. J, Langridge R., and Ferrin T. E. (1982) A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **161**, 269–288.
53. Wlodawer A. and Vondrasek J. (1998) Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct* **27**, 249–284.
54. Wong C. F. and McCammon J. A. (2003) Protein flexibility and computer sided drug design. *Annual Rev Pharmacol Toxicol* **43**, 31–45.
55. Taylor R. D., Jewsbury P. J. and Essex J. W. (2002) A review of protein-small molecule docking methods. *J Comput Aided Mol Des* **16**, 151–166.
56. Muegge I. (2003) Selection criteria for drug-like compounds. *Med Res Rev* **23**, 302–321.
57. Cho S. J., Zheng W., and Tropsha A. (1998) Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J Chem Inf Comput Sci* **38**, 259–268.
58. Jones G., Willett P., Glen R. C., Leach A. R., and Taylor R. (1997) Development and validation of a genetic algroithm for flexible docking. *J Mol Biol* **267**, 727–748.
59. Warren G. L., Andrews C. W., Capelli A. M., Clarke B., LaLonde J., Lambert M. H. et al. (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* **49**, 5912–5931.
60. Keiser M. J., Roth B. L., Armbruster B. N., Ernsberger P., Irwin J. J., and Shoichet B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **25**, 197–206.
61. Keiser M. J., Setola V., Irwin J. J., Laggner C., Abbas A. I., Hufeisen S. J. et al. (2009) Predicting new molecular targets for known drugs. *Nature* **462**, 175–181.
62. Tropsha A. (2005) Application of predictive QSAR models to database mining. In: Oprea T. (Ed.). Cheminformatics in Drug Discovery. Wiley-VCH, Weinheim, 437–455.
63. Tropsha A. and Golbraikh A. (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des* **13**, 3494–3504.
64. Golbraikh A. and Tropsha A. (2002) Beware of q2! *J Mol Graph Model* **20**, 269–276.
65. Tropsha A., Gramatica P., and Gombar V. K. (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *Quant Struct Act Relat Comb Sci* **22**, 69–77.
66. Sachs, L. (1984) Applied Statistics: A Handbook of Techniques. 2nd ed, New York: Springer-Verlag.
67. Golbraikh A., Shen M., Xiao Z., Xiao Y. D., Lee K. H, and Tropsha A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* **17**, 241–253.
68. Irwin J. J. and Shoichet B. K. (2005) ZINC – a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177–182.
69. Medina-Franco J. L., Golbraikh A., Oloff S., Castillo R., and Tropsha A. (2005) Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J Comput Aided Mol Des* **19**, 229–242.
70. Oloff S., Mailman R. B., and Tropsha A. (2005) Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J Med Chem* **48**, 7322–7332.
71. Shen M., Beguin C., Golbraikh A., Stables J. P., Kohn H., and Tropsha A. (2004) Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J Med Chem* **47**, 2356–2364.
72. Zhang S., Wei L., Bastow K., Zheng W., Brossi A., Lee K. H. et al. (2007) Antitumor agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J Comput Aided Mol Des* **21**, 97–112.
73. Tang H., Wang X. S., Huang X. P., Roth B. L., Butler K. V., Kozikowski A. P. et al. (2009)

Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J Chem Inf Model* **49**, 461–476.
74. Hoops S., Sahle S., Gauges R., Lee C., Pahle J., Simus N. et al. (2006) COPASI – a COmplex PAthway SImulator. *Bioinformatics* **22**, 3067–3074.
75. Loew L. M. and Schaff J. C. (2001) The virtual cell: a software environment for computational cell biology. *Trends Biotechnol* **19**, 401–406.
76. Slepoy A., Thompson A. P., and Plimpton S. J. (2008) A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks. *J Chem Phys* **128**, 205101.
77. Salis H., Sotiropoulos V., and Kaznessis Y. N. (2006) Multiscale Hy3S: hybrid stochastic simulation for supercomputers. *BMC Bioinformatics* **7**, 93.
78. Tomita M., Hashimoto K., Takahashi K., Shimizu T. S., Matsuzaki Y., Miyoshi F. et al. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84.
79. Yang C. R., Shapiro B. E., Mjolsness E. D., and Hatfield G. W. (2005) An enzyme mechanism language for the mathematical modeling of metabolic pathways. *Bioinformatics* **21**, 774–780.
80. May E. E. and Schiek R. L. (2009) BioXyce: an engineering platform for the study of cellular systems. *IET Syst Biol* **3**, 77–89.
81. Schiek R. L. and May E. E. (2006) Xyce Parallel Electronic Simulator: Biological Pathway Modeling and Simulation. Albuquerque, NM, Sandia National Laboratories, Report No. SAND2006-1993p.
82. Eldred M. S., Adams B. M., Haskell K., Bohnhoff W. J., Eddy J. P., Gay D. M. et al. (2008) DAKOTA: A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis. Version 4.2 Users Manual. SAND2006-6337.
83. Oishi M. and May E. E. (2007) Addressing biological circuit simulation accuracy: Reachability for parameter identification and initial conditions. Bethesda, MD. IEEE/NIH BISTI Life Science Systems and Application Workshop.
84. Chang A., Scheer M., Grote A., Schomburg I., and Schomburg D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res* **37**(Database issue), D588–D592.
85. Schomburg I., Chang A., Ebeling C., Gremse M., Heldt C., Huhn G. et al. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* **32**(Database issue), D431–D433.
86. Wayne L. G. and Sohaskey C. D. (2001) Nonreplicating persistence of *Mycobacterium tuberculosis*. *Annu Rev Microbiol* **55**, 139–163.
87. Wayne L. G. and Lin K. Y. (1982) Glyoxylate metabolism and adaptation of *Mycobacterium tuberculosis* to survival under anaerobic conditions. *Infect Immun* **37**, 1042–1049.
88. Goodford P. J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **28**, 849–857.
89. Mdluli K. and Spigelman M. (2006) Novel targets for tuberculosis drug discovery. *Curr Opin Pharmacol* **6**, 459–467.
90. Ariens E. J., Simonis A. M., and De Groot W. M. (1955) Affinity and intrinsic-activity in the theory of competitive- and non-competitive inhibition and an analysis of some forms of dualism in action. *Arch Int Pharmacodyn Ther* **100**, 298–322.
91. Olah M., Oprea T. I. (2006) Bioactivity Databases. In: J. B. Taylor, D. J. Triggle (Eds.). Comprehensive Medicinal Chemistry II Vol. **3**. Elsevier, Oxford, 293–313.

Chapter 19

Ligand-Based Approaches to In Silico Pharmacology

David Vidal, Ricard Garcia-Serna, and Jordi Mestres

Abstract

The development of computational methods that can estimate the various pharmacodynamic and pharmacokinetic parameters that characterise the interaction of drugs with biological systems has been a highly pursued objective over the last 50 years. Among all, methods based on ligand information have emerged as simple, yet highly efficient, approaches to in silico pharmacology. With the recent impact on the identification of new targets for known drugs, they are again the focus of attention in chemical biology and drug discovery.

Key words: Target profiling, Polypharmacology, Drug repurposing, Topological descriptors, Structure–activity relationships

1. Introduction

The term in silico pharmacology can be defined as the study of drugs by computational means. In its widest sense, this definition includes developing models and running simulations to acquire knowledge not only on the mode of action of drugs but also on the toxicity, side effects and pharmacokinetic events associated with them when interacting with a biological system [1, 2]. The infancy of in silico pharmacology can be traced back to the early 1960s when correlations between chemical structure and biological activity began to be investigated with the help of the first computers [3]. The field developed into what came to be known as quantitative structure–activity relationships (QSAR) that was quickly established as an essential component of medicinal chemistry and pharmacology [4]. Over the last 50 years, these efforts have generated a plethora of QSAR models on various biological effects (such as binding affinity, absorption, metabolism and toxicity) as well as physico-chemical properties (for example,

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_19, © Springer Science+Business Media, LLC 2011

partition coefficient, aqueous solubility, vapour pressure and dissociation constant), many of which have been collected and stored in the C-QSAR database [5].

In the late 1990s, with computer-based methods having reached sufficient maturity, the concept of in silico screening emerged in an attempt to complement existing experimental high-throughput screening (HTS) techniques that had disappointingly poor performances at costs much higher than originally expected [6]. Up until then, QSAR had traditionally focused on small sets of congeneric compounds. From then on, in silico screening has extended the applicability of computational models along the chemical dimension defined by all molecules, both synthesised and plausible of being synthesised. Under this new scheme, molecules in large chemical libraries can be first filtered based on certain pre-defined drug-like properties and subsequently scored and ranked according to their likelihood of having affinity for a certain target [7]. From a ligand-based perspective, this likelihood is assessed according to the similarity to one or multiple known bioactive ligands [8]. All molecules in a particular database are then ranked by the similarity to reflect decreasing probability of being active and the top-scoring ones can be prioritised for going into experimental testing, thus representing a knowledge-based cost-effective strategy in drug discovery [9]. Over the years, the pharmaceutical industry has learnt to accept that in silico screening methods can be an efficient counterpart to HTS to the point that they have undoubtedly become an integral part of today's lead generation process [10, 11].

In recent years, in silico screening has been extended further along the biological dimension leading to new integrative methods capable of estimating the pharmacological profile of molecules on multiple targets [12]. The development of in silico profiling methods based on ligand information has been possible due to multiple coordinated efforts to generate, compile and store pharmacological data of compounds [13]. Table 1 collects some of the current public and commercial resources that contain data on ligand–protein interactions, some of them being offered through web-based user-friendly environments with searching and browsing capabilities.

All these data are now available to be exploited for the development of computational systems that annotate small molecules to protein targets based on their similarity to known bioactive compounds. These methods currently have a strong impact in drug discovery as a means for detecting affinities for additional targets other than the primary target(s), which may lead to the identification of new therapeutic opportunities as well as to the anticipation of potential side effects and toxicities [14, 15].

In the remainder of this chapter, focus will be given, first, to the molecular descriptors and, then, to the ligand-based

Table 1
List of public and commercial resources of pharmacological data for small molecule ligands

Resource	Website
DrugBank	http://www.drugbank.ca
IUPHARdb	http://www.iuphar-db.org
BindingDB	http://www.bindingdb.org
BindingMOAD	http://www.bindingmoad.org
PDSP	http://pdsp.med.unc.edu
ChEMBLdb	http://www.ebi.ac.uk/chembldb/
AffinDB	http://pc1664.pharmazie.uni-marburg.de/affinity/
PubChem	http://pubchem.ncbi.nlm.nih.gov
CTD	http://ctd.mdibl.org
STITCH	http://stitch.embl.de
PharmGKB	http://www.pharmgkb.org
KEGG	http://www.genome.jp/kegg/
GLIDA	http://pharminfo.pharm.kyoto-u.ac.jp/services/glida/
NikkajiWeb	http://nikkajiweb.jst.go.jp
BioPrint	http://www.eidogen-sertanty.com/products_kinasekb.html
WOMBAT	http://www.sunsetmolecular.com
GVKBio	http://www.gvkbio.com/informatics.html
Aureus	http://www.aureus-pharma.com/Pages/Products/Aurscope.php
Symyx	http://www.symyx.com/products/databases/bioactivity/
Eidogen-Sertanty	http://www.eidogen-sertanty.com/products_kinasekb.html

approaches to in silico pharmacology developed in our laboratory, with a final discussion devoted to the latest applications aiming at identifying novel targets for known drugs that may lead to novel therapeutic opportunities or explain some of their recognised side effects.

2. Molecular Descriptors

A key aspect to all ligand-based approaches to in silico pharmacology is the use of appropriate molecular descriptors to represent chemical structures mathematically, as they will determine to

which extent molecules are perceived similar. As a result, a large number and variety of molecular descriptors reflecting the one-dimensional, two-dimensional (2D) and three-dimensional features of chemical structures have been developed [16, 17] and procedures to select those that are most relevant for modelling the biological effect of interest have been devised [18]. Among them, 2D molecular fingerprints remain to be the most popular molecular descriptors. They are composed of bit strings that encode the absence (0) or presence (1) of some 2D atom paths, fragments or substructures present in a predefined dictionary. Frequently used structural fingerprints are MACCS structural keys (MDL Information Systems), Daylight (DAYLIGHT), Unity (Tripos) and ECFP4 fingerprints [19].

In our efforts to develop an integrated computational system to in silico pharmacology, we have implemented three new types of molecular descriptors that will be described in detail in the following sections. They have been named as PHRAGs, FPDs and SHED.

2.1. Pharmacophoric Fragments

A Pharmacophoric Fragment (PHRAG) is a pharmacophoric-based descriptor derived from a direct fragmentation of the molecular graph into overlapping segments of a fixed length. Figure 1 illustrates in detail how PHRAGs are derived for salicylic acid (only non-hydrogen atoms are considered). The selection of the most suitable PHRAG length is critical for the success of the method, and it may vary depending on the specific problem. However, in order to ensure their applicability for in silico screening and profiling purposes, the optimised PHRAG length must reflect the correct balance between the encoded information,

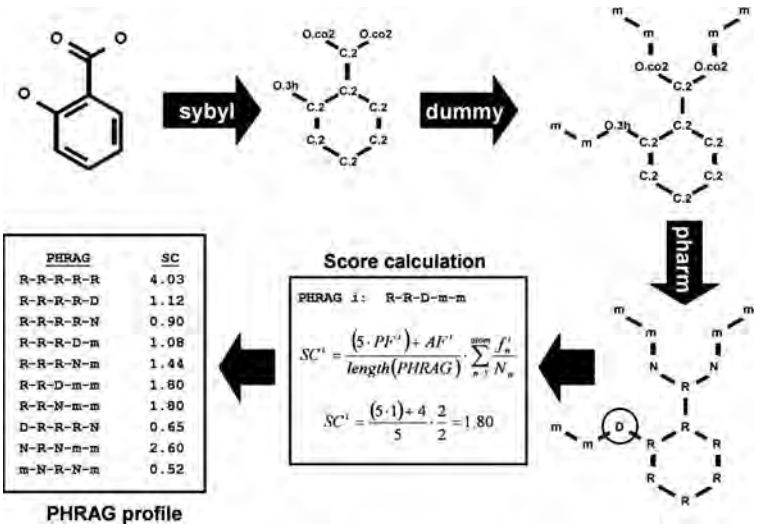


Fig. 1. Generation of a PHRAG profile from salicylic acid.

which increases with length, and the requirement of a statistically acceptable representation of PHRAGs in different molecules, which decreases with length. In our implementation, a fixed length of five atom-centred pharmacophoric features per fragment has been selected.

The process of extracting the PHRAG profile from a molecule is as follows: first, every atom in the molecule is mapped onto a Sybyl atom type and assigned to one of seven atom features, namely, hydrophobic (H), aromatic (R), donor (D), acceptor (A), acceptor–donor (W), positively charged (P) or negatively charged (N); then, the entire molecular graph is analysed and all PHRAGs present in the molecule are extracted and assigned to the central atom of the fragment. In order to allow equal representation of all atoms in the molecule and avoid misrepresenting terminal atoms, a number of dummy atoms (assigned to Sybyl atom type "m") are added to every terminal atom to fulfil the PHRAG length requirement (Fig. 1). In addition, one must take into consideration that highly branched atoms and ring systems tend to generate a large number of PHRAGs. Therefore, in order to correct for this bias, a score value (SC^i) of every unique PHRAG i accounting for the branching level is calculated according the Eq. 1:

$$SC^i = \frac{\left(5 \cdot PF^i\right) + AF^i}{\text{length(PHRAG)}} \cdot \sum_{n=1}^{\text{atom}} \frac{f_n^i}{N_n} \tag{1}$$

where PF^i and AF^i are the number of polar and apolar features in PHRAG i, N_n is the total number of PHRAGs assigned to central atom n and $f_n{}^i$ is the number of occurrences of PHRAG i assigned to central atom n. The score includes also a correction for the misrepresentation of polar PHRAGs in molecules. This correction was extracted from a previous analysis on over hundred thousand bioactive molecules, in which on average polar and apolar features were found to follow a 1:5 ratio.

The similarity between the PHRAG profiles of two molecules A and B can be computed using Eq. 2. In this equation, $SC_A{}^i$ is the score of PHRAGs i in molecule A, $SC_B{}^i$ is the score of PHRAGs i in B and n_T is the number of unique PHRAGs contained in either molecule A or B. Molecules having the same types and scores of PHRAGs will achieve a similarity of 1:

$$\text{PHRAGsim}^{A,B} = 1 - \frac{\sum_{i=1}^{n_T} \left|SC_A^i - SC_B^i\right|}{\sum_{i=1}^{n_T} SC_A^i + \sum_{i=1}^{n_T} SC_B^i} \tag{2}$$

To illustrate the type of hits identified by means of PHRAGsim, the catalogue of all chemical suppliers was screened against salicylic acid. A sample of hits ordered on the basis of PHRAGsim relative to salicylic acid is provided in Fig. 2. As can be observed,

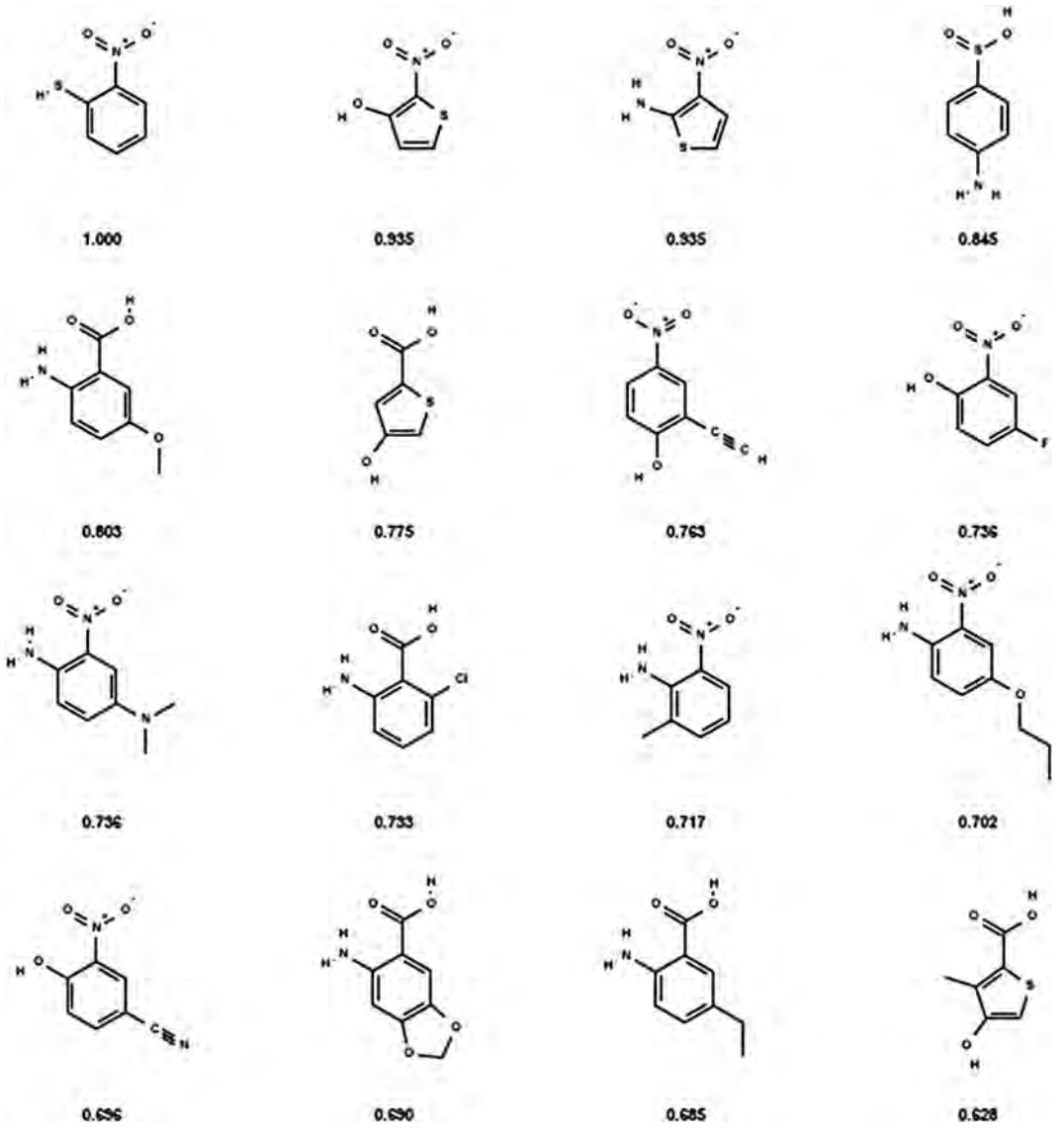


Fig. 2. Example of retrieved hits in descending order based on PHRAGsim values against salicylic acid.

they mostly retain the essential pharmacophoric fragments present in the reference compound, with increasing variety being allowed as the similarity value decreases.

2.2. Feature-Pair Distributions

Feature-Pair Distributions (FPDs) belong to the family of atom-pair molecular descriptors, in which all pairs of atoms in a molecule are encoded according to the shortest bond path connecting them. For each feature pair, the overall distribution captures the occurrence of that particular pair of features at different predefined bond-path distance ranges. A key step in the process is the mapping of every non-hydrogen atom in a molecule to a Sybyl

atom type. Subsequently, each atom type is assigned to one or more of six pharmacophoric features, namely, hydrophobic (H), aromatic (R), donor (D), acceptor (A), positively charged (P) and negatively charged (N). Then, the shortest bond-path length between atom-centred feature pairs is extracted, and its occurrence at different path lengths is stored in the corresponding 21 feature-pair distributions emerging from the combination of all possible pairs of pharmacophoric features (HH, HR, HA, HD, HN, HP, RR, RA, RD, RN, RP, AA, AD, AN, AP, DD, DN, DP, NN, NP, PP). In our current implementation, distances of up to 20 bonds are considered and, for every pair, ten distance bins are used corresponding to a distance length of 1, 2, 3, 4, 5, 6, 7–9, 10–12, 13–15, 16–20 bonds between the atom-centred features. Feature pairs at distances over 20 bonds are ignored.

Prior to evaluating the similarity between two FPDs, pharmacophoric features are classified as either polar (A, D, N, P) or apolar (H, R), leading to three classes of FPD comparisons: apolar-apolar ("aa"), apolar-polar ("ap") and polar-polar ("pp"). For every class pair, a similarity measure is defined based on the overlap of the corresponding feature-pair distributions. The final FPD similarity is defined as the average value among all pair classes ("aa", "ap" and "pp"), present in either of the molecules being compared, according to Eq. 3:

$$\mathrm{FPDsim}^{A,B,\mathrm{CLASS}} = \frac{\sum_{i=1}^{\mathrm{BINS}} \min(\mathrm{POP}_A^i, \mathrm{POP}_B^i)}{\sum_{i=1}^{\mathrm{BINS}} \mathrm{POP}_A^i + \sum_{i=1}^{\mathrm{BINS}} \mathrm{POP}_B^i} \tag{3}$$

where $\mathrm{POP}_A{}^i$ is the population of bin i in molecule A, $\mathrm{POP}_B{}^i$ is the population of bin i in B and BINS is the total number of bins from all the feature-pair distributions included in pair class CLASS and occupied by either molecule A or B. For the sake of clarity, the FPD similarity between chlorpromazine and imipramine is shown in Fig. 3.

2.3. Shannon Entropy Descriptors

Shannon entropy descriptors (SHEDs) are derived from distributions of atom-centred feature pairs extracted directly from the topology of molecules [20]. The process of obtaining a SHED from a chemical structure requires deriving the corresponding FPD first, as described above. However, the slight difference here is that, in order to speed up calculations with SHEDs, only four atom-centred features are considered, namely, hydrophobic (H), aromatic (R), acceptor (A) and donor (D), which in turn leads to 10 FPDs instead of 21 (HH, HR, HA, HD, RR, RA, RD, AA, AD, DD). Once the FPD has been obtained, the concept of Shannon entropy is applied to determine the variability of the bin population along each feature-pair distribution. Within this

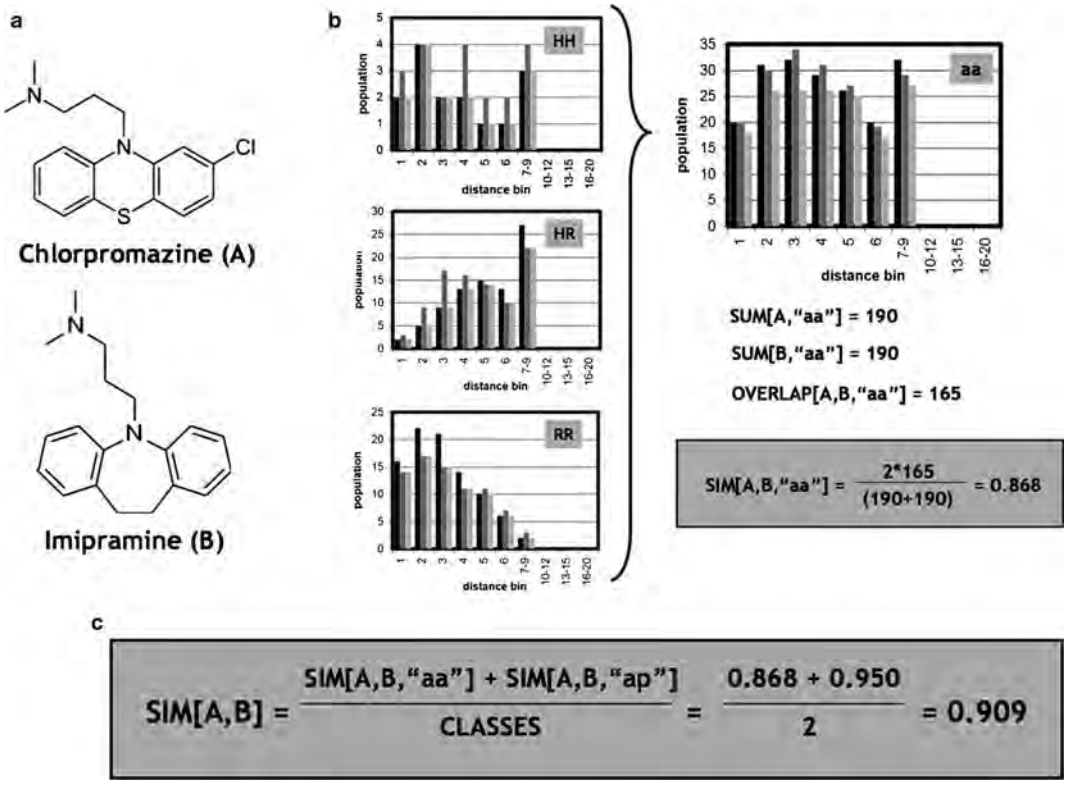


Fig. 3. The process of calculating the similarity between the FPDs derived from (**a**) the structures of chlorpromazine and imipramine. (**b**) Calculation of the apolar–apolar FPD similarity; bin populations from chlorpromazine and imipramine are depicted with *black* and *grey bars*, respectively, and the corresponding overlap between them by *light grey bars*. (**c**) Obtaining the final FPD similarity value.

approach, the entropy, S, of a population, P, distributed in a certain number of bins (representing in this case the different path lengths), $N = 20$, is given by Eq. 4:

$$S = -\sum_{i=1}^{N} \rho_i \ln \rho_i; \qquad \rho_i = p_i/P \tag{4}$$

where ρ_i and p_i are the probability and the population, respectively, at each bin i of the distribution. The values of S range between 0, reflecting the situation of the entire population being concentrated in a single bin, and a maximum number, $S_{\max} = \ln N$, reflecting the situation of a uniformly distributed population among all bins. To have a more intuitive measure that can be linearly related to the situation of full uniform occupancy, entropy values are transformed into projected entropy values, $E = e^S$. Accordingly, E values provide a measure of the expected maximum uniform occupancy from the corresponding S value. Now, for any given population $P > 0$, the values of E can vary from 1, reflecting the situation of zero entropy in which the population is totally

concentrated in a single bin, to N, reflecting the situation of maximum entropy in which the population is uniformly distributed among all bins. In the limit case of $P = 0$, then E will be assigned to $E = 0$. This E value will ultimately be the Shannon entropy descriptor (SHED) of the corresponding feature pair. The set of SHED values obtained for the ten possible feature pairs constitutes the SHED profile of a molecule.

3. Recent Applications

An arsenal of computational methods for in silico pharmacology has been developed over the years, and they currently are applicable to almost any aspect of relevance to pharmacology. In particular, an explosion of ligand-based approaches to in silico target profiling has been observed lately, which have lead to some top stories on the identification of new targets for known drugs. Accordingly, emphasis will be put on this latter aspect in the following sections.

3.1. High-Throughput QSAR

Traditionally, deriving QSAR models is usually associated with a computational activity focussed on a single target for which biological affinities for a limited number of congeneric compounds are available. The recent compilation of pharmacological data for hundreds of thousands of compounds on thousands of protein targets is now offering the possibility to derive QSAR models in a large scale. Taking a combination of both public (DrugBank and BindingDB) and commercial (WOMBAT) annotated chemical libraries, we developed a strategy for the high-throughput generation of QSAR models. PHRAG descriptors were used to represent molecules mathematically. The prediction of binding affinities was performed by nearest-neighbour interpolation, in which the predicted affinity is the result of averaging the affinities from all compounds with PHRAGsim values higher than a given threshold (0.75 in this case). In total, 1,924 QSAR models were generated, covering 1,121 targets, namely, 677 enzymes (including 36 cytochromes P450 and 214 kinases), 189 G protein-coupled receptors (GPCRs), 113 ion channels and transporters and 29 nuclear receptors (Fig. 4). The average leave-one-out correlation coefficient for all QSAR models derived from the validation set is $q^2 = 0.73$, with a root mean square error of qRMS = 0.74 and over 80% of molecules having predicted affinity values within 1 log unit of their corresponding experimental value. These QSAR models can now be routinely applied to predict the binding affinity on 1,121 protein targets for compounds within the established applicability domain of each model (defined by the PHRAGsim threshold provided above).

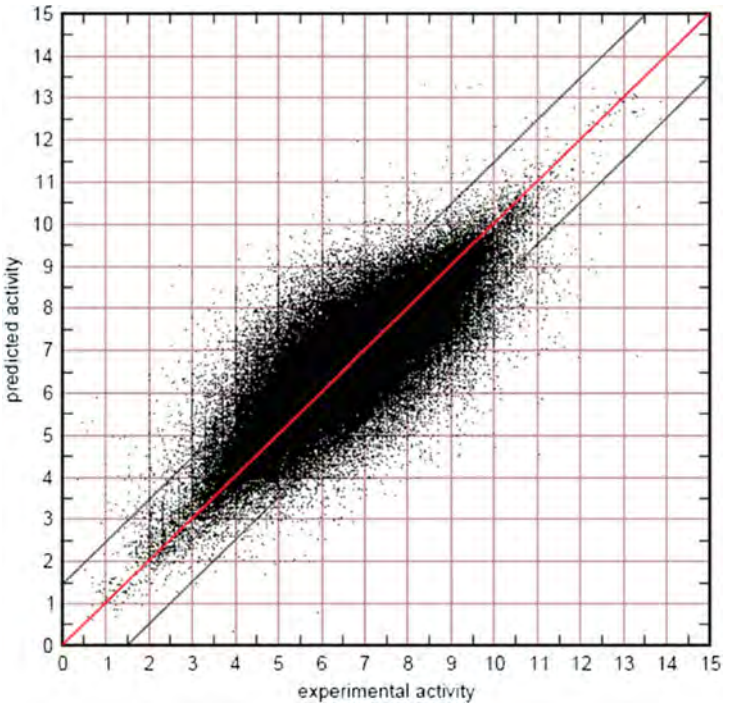


Fig. 4. Scatter plot of the experimental vs. predicted activities for 1,924 QSAR models covering 1,121 biological targets ($q^2 = 0.73$; qRMS = 0.74).

3.2. In Silico Target Profiling

The performance of our current ligand-based approach to in silico target profiling was assessed against a set of 790 drugs from DrugBank for which a total of 3,775 interactions (those being pK_i, pK_d, pIC_{50}, or $pEC_{50} \geq 5$) with protein targets could be retrieved from all public sources (Table 1). The final prediction of the drug–target binding affinities is calculated by averaging the predicted affinities obtained by interpolation from each one of the three descriptors implemented internally (PHRAGs, FPDs and SHED). The results are graphically summarised in Fig. 5.

As can be observed, we can correctly retrieve 48.7% of the interactions in the value range of [5,6), which represent 34.9% of all interactions considered. The recall increases significantly as one moves to the bins related to stronger interactions. For example, we retrieve 65.8 and 74.3% of the interactions in the value ranges of [6,7) and [7,8), respectively, which represent 25.5 and 21.0% of all interactions under 10 μM. Finally, the percentage of interactions recovered for values in the range of [8,9) and >9 are both above 80%, although they collectively represent only 18.6% of all interactions.

The present results are highly encouraging towards the potential applicability of these methods to identify new targets for known drugs (as will be shown in the next section). As a final

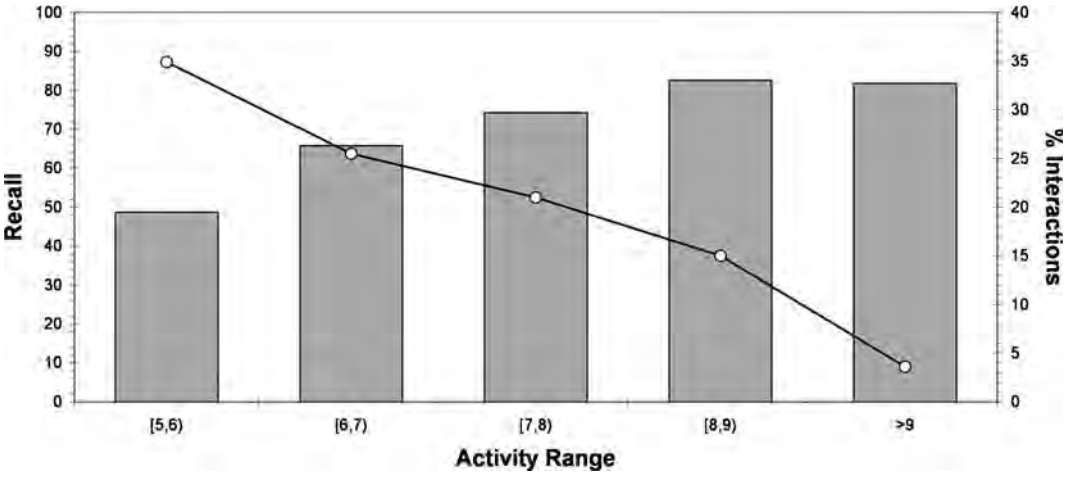


Fig. 5. Percentage of known drug–target interactions retrieved (recall) within each range of activity values (*grey bars*) and representative percentage of the interactions within each activity range from all known interaction under 10 μM.

remark, however, a close inspection to the distribution of all drug–target interactions considered above among the different protein family reveals that over 52% (60%) of all interactions under 10 μM (1 μM) are linked to a GPCR target, a protein family recognised by its relatively high internal degree of cross-pharmacology [20].

3.3. Identification of New Targets for Known Drugs

Using SHED descriptors, Gregori-Puigjané and Mestres [21] were able to validate their in silico target profiling approach by recovering affinities for the α_{1A} adrenergic receptor on a set of drugs pharmacophorically similar to known α_{1A} drugs but with essentially different scaffolds. The same approach was later applied to profiling a set of 767 drugs on 684 targets [22]. Analysis of the most connected part of the target network revealed that aminergic GPCRs were highly connected to opioid, sigma and NMDA receptors, which anticipated drug polypharmacologies among these receptors. In particular, opioid receptors were identified as potential new targets for pergolide, interactions that were later confirmed experimentally for the κ and μ opioid receptors (Scott Boyer, personal communication).

A set of proteins (or protein families) sharing similar bioactive compounds is said to be related by cross-pharmacology. In this context, Keiser et al. [23] were able to identify cross-pharmacologies between several target pairs by applying a statistical approach based on cumulative similarity between sets of bioactive ligands (SEA). Among them, the enzymes thymidylate synthase and dihydrofolate reductase were found to have related pharmacologies despite having no substantial sequence

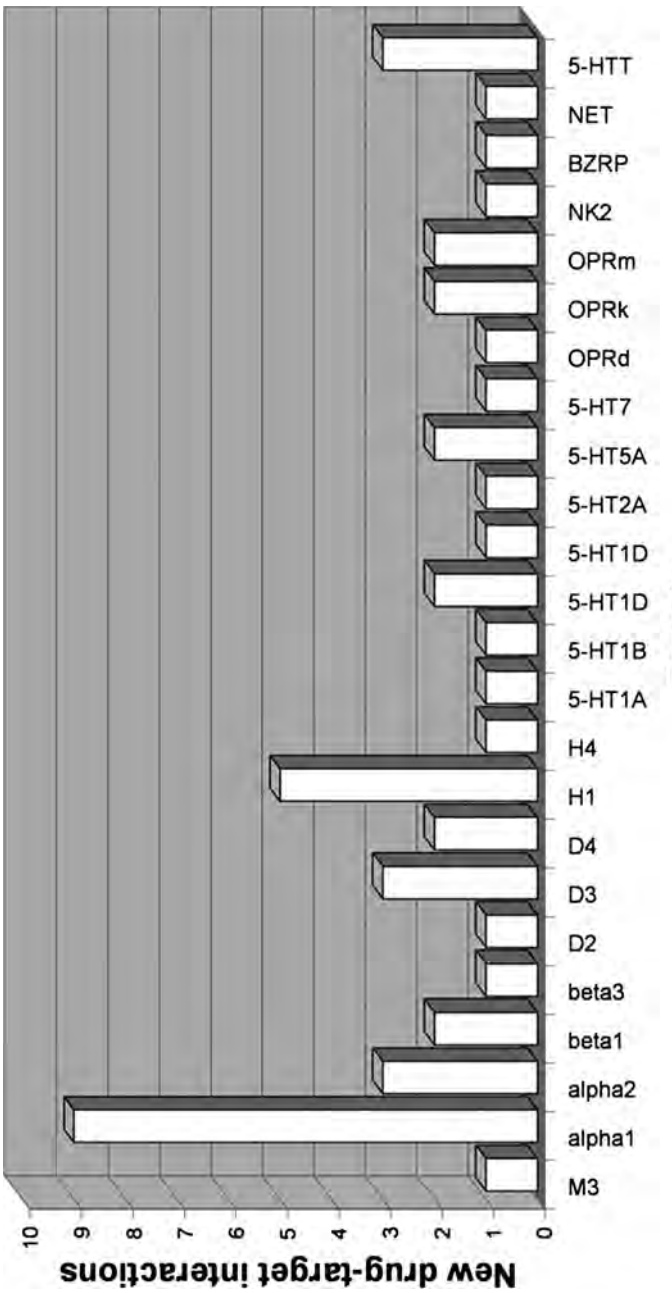


Fig. 6. The distribution of all 48 new drug–target interactions collectively reported in recent works [22–25].

identity and being structurally distinct. The identification of cross-pharmacology relationships among targets leads naturally to the exploitation of those target connections for the prediction of novel targets for drugs with already known target(s). Under this premise, unreported links between methadone, emetine, and loperamide and muscarinic M3, α2 adrenergic, and neurokinin NK2 receptors, respectively, were identified [23]. More recently, Keiser et al. [24] provided experimental confirmation of 23 new drug–target associations predicted with their similarity-based SEA approach.

Instead of using similarity to ligands or ligand sets, Campillos et al. [25] used side-effect similarity to identify drugs that agglomerate from a phenotypic viewpoint. Under this approach, they reported 13 novel drug–target relations, mostly involving aminergic GPCRs, namely, histamine H1, dopamine D3 and serotonin 5-HT_{1D} receptors.

An analysis of the complete list of 48 new drug–target interactions reported by all works commented above reveals that 43 are interactions to GPCRs, of which 37 are aminergic GPCRs, the majority of them predicted for drugs already known to bind to other GPCRs already (Fig. 6). Therefore, a challenge for the coming years will be to identify new targets beyond close phylogenetic relationships, that is, beyond statistically significant cross-pharmacologies.

4. Conclusion and Outlook

The recent applications of ligand-based approaches to in silico pharmacology have provided ample evidence of the key impact that these rather simple, yet highly efficient, computational methods have in both chemical biology and drug discovery. With further developments and wider validation benchmarks, a stage of maturity is foreseeable in the coming years, implying that the study of drugs by computational means will gradually ascend towards completeness.

Acknowledgments

Funding for this research was received from the Instituto de Salud Carlos III and the Spanish Ministerio de Ciencia e Innovación (project BIO2008-02329). GRIB is a node of the Instituto Nacional de Bioinformática (INB) and a member of the RETIC COMBIOMED network.

References

1. Ekins, S., Mestres, J., and Testa, B. (2008) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *Br. J. Pharmacol.* **152**, 9–20.
2. Ekins, S., Mestres, J., and Testa, B. (2008) In silico pharmacology for drug discovery: applications to targets and beyond. *Br. J. Pharmacol.* **152**, 21–37.
3. Hansch, C. and Fujita, T. (1964) Rho-sigma-pi analysis: a method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **86**, 1616–1626.
4. Hansch, C., Hoekman, D., Leo, A., Weininger, D., and Selassie, C. D. (2002) Chembioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem. Rev.* **102**, 783–812.
5. Kurup, A. (2003) C-QSAR: a database of 18,000 QSARs and associated biological and physical data. *J. Comput. Aided Mol. Des.* **17**, 187–196.
6. Lahana, R. (1999) How many leads from HTS? *Drug Discov. Today* **4**, 447–448.
7. Bajorath, J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **1**, 882–894.
8. Willett, P. (2003) Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **31**, 603–606.
9. Lengauer, T., Lemmen, C., Rarey, M., and Zimmermann, M. (2004) Novel technologies for virtual screening. *Drug Discov. Today* **9**, 27–34.
10. Bleicher, K. H., Böhm, H. -J., Müller, K., and Alanine, A. I. (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2**, 369–378.
11. Shoichet, B. K. (2004) Virtual screening of chemical libraries. *Nature* **432**, 862–865.
12. Mestres, J. (2004) Computational chemogenomic approaches to systematic knowledge-based drug discovery. *Curr. Top. Drug Discov. Dev.* **7**, 304–313.
13. Savchuk, N. P., Balakin, K. V., and Tkachenko, S. E. (2004) Exploring the chemogenomic knowledge space with annotated chemical libraries. *Curr. Opin. Chem. Biol.* **8**, 412–417.
14. Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genetics* **5**, 262–275.
15. Bajorath, J. (2008) Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **12**, 352–358.
16. Karelson, M. (2000) Molecular descriptors in QSAR/QSPR. Wiley-VCH: New York.
17. Todeschini, R. and Consonni, V. (2000) Handbook of molecular descriptors. Wiley-VCH: New York.
18. Walters, W. P. and Goldman, B. B. (2005) Feature selection in quantitative structure-activity relationships. *Curr. Opin. Drug Discov. Devel.* **8**, 329–333.
19. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053.
20. Mestres, J., Gregori-Puigjané, E., Valverde, S., and Solé, R. V. (2009) The topology of drug-target interaction networks: implicit dependence on drug properties and protein families. *Mol. Biosyst.* **5**, 1051–1057.
21. Gregori-Puigjané, E. and Mestres, J. (2006) SHED: Shannon entropy descriptors from topological feature distributions. *J. Chem. Inf. Model.* **46**, 1615–1622.
22. Gregori-Puigjané, E. and Mestres, J. (2008) A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High Throughput Screen.* **11**, 669–676.
23. Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. **25**, 197–206.
24. Keiser, M. J., Setola, V., Irwin, J. J., Laggner, C., Abbas, A. I., Hufeisen, S. J., Jensen, N. H., Kuijer, M. B., Matos, R. C., Tran, T. B., Whaley, R., Glennon, R. A., Hert, J., Thomas, K. L. H., Edwards, D. D., Shoichet, B. K., and Roth, B. L. (2009) Predicting new molecular targets for known drugs. *Nature* **462**, 175–182.
25. Campillos, M., Kuhn, M., Gavin, A. -C., Jensen, L. J., and Bork, P. (2008) Drug target identification using side-effect similarity. *Science* **321**, 263–266.

Chapter 20

Molecular Test Systems for Computational Selectivity Studies and Systematic Analysis of Compound Selectivity Profiles

Dagmar Stumpfe, Eugen Lounkine, and Jürgen Bajorath

Abstract

For chemical genetics and chemical biology, an important task is the identification of small molecules that are selective against individual targets and can be used as molecular probes for specific biological functions. To aid in the development of computational methods for selectivity analysis, molecular benchmark systems have been developed that capture compound selectivity data for pairs of targets. These molecular test systems are utilized for "selectivity searching" and the analysis of structure–selectivity relationships. Going beyond binary selectivity sets focusing on target pairs, a methodological framework, Molecular Formal Concept Analysis (MolFCA), is described for the definition and systematic mining of compound selectivity profiles.

Key words: Chemical biology, Compound selectivity, Structure–selectivity relationships, Selectivity searching, Formal concept analysis, Molecular formal concept analysis

1. Introduction

In recent years, highly interdisciplinary and in part overlapping research fields have evolved at the interface between chemistry and biology including *chemical biology*, *chemogenomics*, and *chemical genetics*. Despite differences in their principal goals and research approaches, these disciplines have in common that they focus on the effects of small molecules on biological systems, for example, by inducing or reverting specific biological phenotypes (*chemical genetics*) or by exploring functions associated with biological systems or processes (*chemical biology*) [1–5]. The boundaries between these disciplines are rather fluid, but *chemogenomics* also has a strong conceptual link to modern drug discovery by aiming to comprehensively study possible drug–target interactions [1, 5, 6].

Among different tasks that are associated with the use of small molecules to explore biological functions is the assessment

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_20, © Springer Science+Business Media, LLC 2011

of compound selectivity. For example, in order to dissect the functional profiles of closely related members of protein families, small molecules are required that are capable of differentiating between such targets, i.e., inhibit or antagonize an individual target with high selectivity over others. Because the identification of target-selective compounds currently largely depends on extensive biological screening and chemical optimization efforts, interest in the computational analysis and prediction of compound selectivity is growing [5]. Given the emerging notion of *polypharmacology* [7, 8], it is becoming increasingly clear that apparent ligand selectivity within a target family does often not result from exclusive binding events, but rather differential binding to multiple targets. Therefore, it is attractive to explore structure–selectivity relationships (SSRs) of small molecules against multiple targets. For this purpose, computational approaches would be very helpful and one would ultimately like to predict target selectivity or selectivity profiles of candidate compounds in order to reduce the magnitude of experimental efforts. A limiting factor for computational selectivity analysis has thus far been the lack of suitable molecular test systems containing compounds with different selectivity profiles that could be used to benchmark computational methods.

In this chapter, we describe initial studies designed to systematically assess compound selectivity by computational means. In order to generate suitable molecular test systems, selectivity has to be defined formally on the basis of potency ratios. We show how binary selectivity data sets are assembled for selectivity searching and the analysis of SSRs. Going beyond binary selectivity sets that focus on target pairs, we also describe an approach for the definition and systematic mining of selectivity profiles termed Molecular Formal Concept Analysis (MolFCA). MolFCA can be used to identify structurally diverse compounds in biologically annotated databases that share defined selectivity profiles of varying complexity.

2. Compound Selectivity

The selectivity of a ligand results from preferential (high-potency) binding to one target within a pair. If members of a protein family have high binding site similarity, multiple binding events are likely to occur. Furthermore, most binding sites display a certain degree of permissiveness to ligand variation, which further increases the potential of cross-target binding. The assembly of compound sets that capture selectivity information arising from multiple binding events requires the analysis of compound potency annotations against multiple targets and the calculation and organization of selectivity values.

2.1. Potency Ratio as a Selectivity Criterion

For the assembly of binary selectivity data sets, compound selectivity is defined on the basis of differences in potency values of a ligand against two target proteins. The potency ratio of a ligand for two targets (T_A, T_B), in the following also referred to as *selectivity ratio* (SR), is calculated from either its K_i or its IC_{50} values as follows:

$$T_A/T_B \equiv SR(ligand, T_A, T_B) = 1 : \frac{K_{i(ligand,T_A)}}{K_{i(ligand,T_B)}}$$

K_i is the dissociation constant of an enzyme-inhibitor complex, and the IC_{50} value represents the concentration of an inhibitor required for 50% inhibition of the enzymatic reaction at a specific substrate concentration. For competitive inhibitors, these different measures are related to each other by the Cheng–Prusoff equation [6].

SR is utilized as a *selectivity criterion* that has to be defined.

2.2. Design of Molecular Selectivity Sets

For the selectivity sets described here, 50-fold higher potency of a compound against target A than B, corresponding to SR(A/B) = 50, is used as a selectivity criterion, and the compound is considered to be *selective* for target A over B. By contrast, a ligand is considered active but nonselective if SR is equal or smaller than 10. Compounds with an SR between 10 and 50 are not utilized in order to avoid boundary effects in selectivity assignments (i.e., compounds with only a marginal difference in their SR, e.g., an SR of 49 and 51, respectively, would be classified differently). In the following, the design of two binary selectivity set systems is discussed [8, 9].

2.2.1. Selective and Inverse-Selective Ligands

The first system contains binary selectivity sets for several biological targets from different protein families [10]. For each target, ligands with experimentally confirmed potency information for at least one additional related target were assembled, and all compounds meeting the SR 50 criterion were added to a target A over B (A/B) set. Ligands with inverse selectivity, i.e., SR (B/A) greater or equal 50, were assembled in a target B over A (B/A) set [10]. The distribution of potency values and application of selectivity criteria are visualized in Fig. 1a that illustrates this process for inhibitors of cathepsin K and S. Each data point represents a ligand, and its coordinates are defined by its potency values against both enzymes. Black dots mark inhibitors that are included in the cathepsin K over S (K/S) and S over K (S/K) selectivity sets. One inhibitor from each set is circled and shown in more detail in Fig. 1b. Both compounds have differential potency for the targets cathepsin K and S: Ligand 1 has a 50-fold lower potency value for cathepsin S than K and is thus considered to be selective for cathepsin S over K (S/K, SR 50) whereas ligand 2 shows target

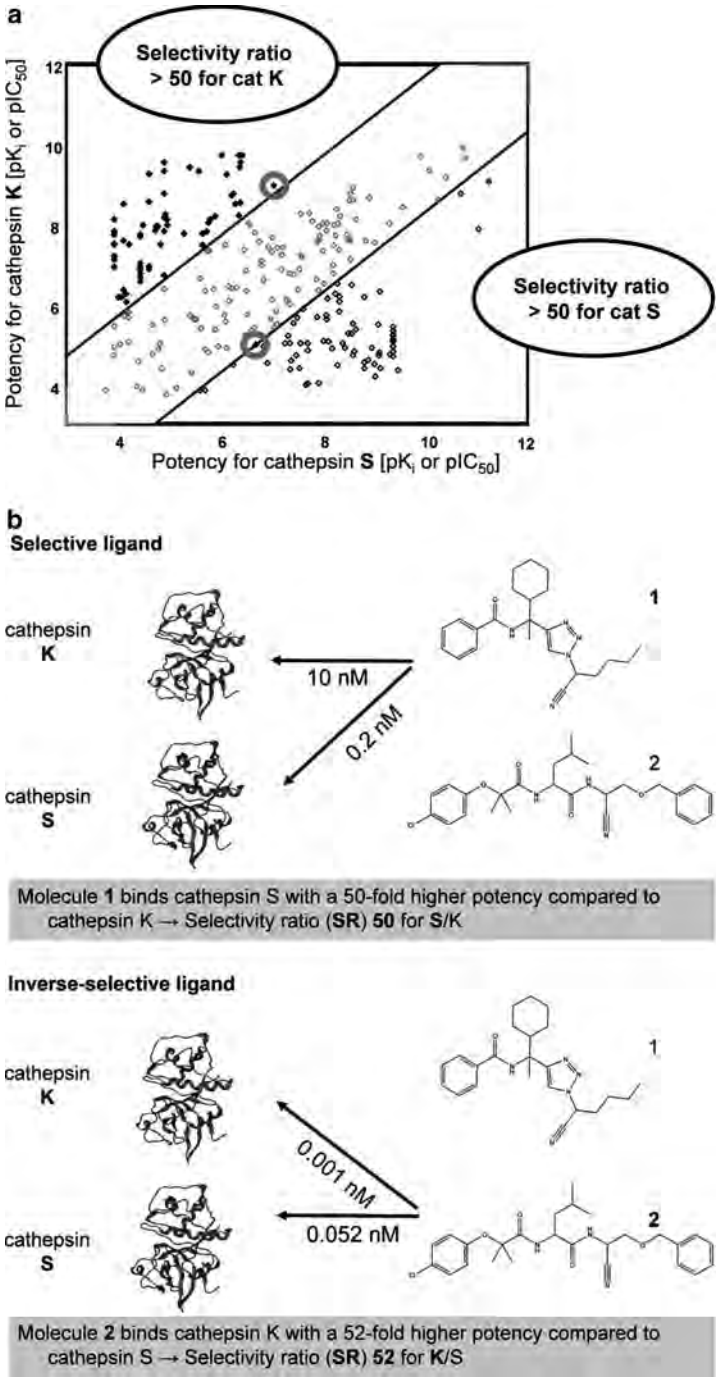


Fig. 1. Classification of selective and inverse-selective compounds. (**a)** Inhibitors of cathepsin (cat) K and S are shown. Potency values against cat S are plotted against the corresponding values for cat K on the pK_i and pIC_{50} scale. Diagonals delineate the 50-fold potency intervals. Compounds selective cat K over A and cat S over K are colored *black* (*filled and unfilled dots*). (**b**) An exemplary molecule from each set is shown and the calculation of selectivity ratio (SR) illustrated.

selectivity for cathepsin K over S (K/S, SR 52). Sets of ligands with inverse selectivity have been used for computational selectivity analyses. SSRs (Subsection 2.3) have been explored, and computational methods for selectivity searching have been evaluated (Subsection 2.4).

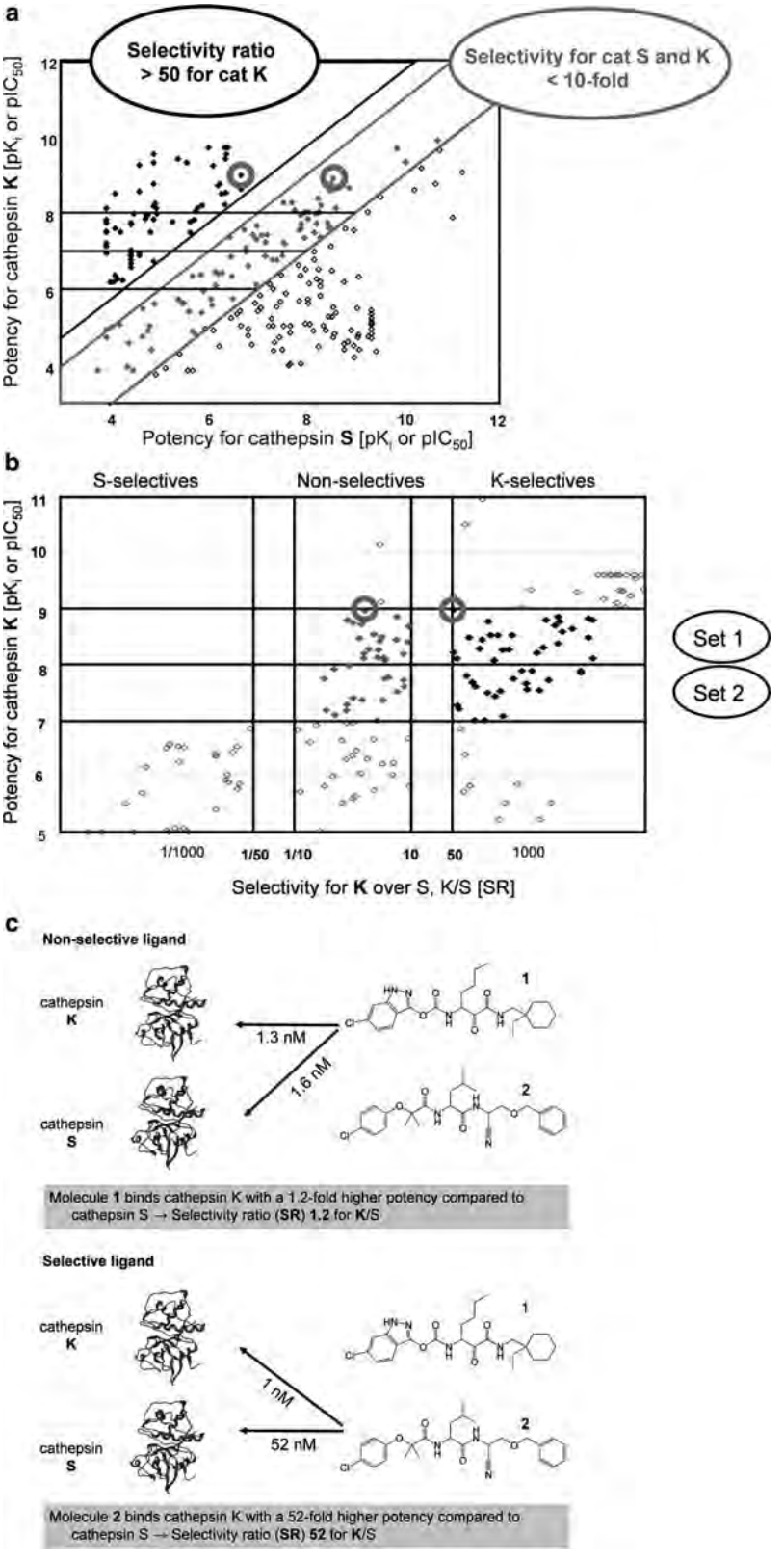
2.2.2. Selective and Nonselective Ligands

The next generation of selectivity sets is more elaborate than the systems discussed above. This is the case because second-generation sets contain ligands that are either selective or nonselective for one target over another [11]. Additionally, selective and nonselective compounds are classified according to different potency levels [11].

Ligand information for different members of four protein families has been collected and ligands organized into nonselective or selective compound sets or omitted if failing to meet the following criteria: Selective ligands must have an SR (A/B) of at least 50 and nonselective ligands an SR (A/B) of 10 or smaller. Thus, nonselective compounds have comparable potency against the two targets, as illustrated in Fig. 2a. Again, inhibitors of cathepsin K and S are shown. Compounds meeting the selectivity (SR 50) and non-selectivity (SR 10) criteria are collected in different sets. Furthermore, each selectivity set contains only ligands that inhibit the target within a defined potency range. Molecules with potency values within one order of magnitude on the pK_i or pIC_{50} scale, i.e., with values in the range of pK_i or pIC_{50} 6–7, or 7–8, or 8–9, are assembled in a different selectivity set (*see* Figs. 2a, b). This design of selectivity sets ensures that structural variations between subsets (i.e., selective and nonselective subset) are directly related to selectivity and are not the result of compounds having very different potency (e.g., hits vs. optimized leads). Also, none of the sets contains weakly potent and highly potent compounds that are selective for a given target. Figure 2c shows a selective and a nonselective ligand of cathepsin K compared to S.

2.3. Structure–Selectivity Relationships

The *similarity property principle* (SPP) states that structurally similar molecules are likely to have similar biological activity [12], and thus global molecular similarity should be an indicator for similar biological activity of small molecules. However, the SPP is not applicable in the case of *discontinuous* and *heterogenous* SARs where minor structural modifications lead to substantial changes in compound potency [13, 14]. This situation is further complicated when biological activity for a second target is taken into account and compound selectivity information considered. In the selectivity sets described above, structural similarity correlates with selectivity. Structural modifications of ligands are likely to lead to different changes in potency against individual targets, which often leads to complex SSR phenotypes. The analysis of the two selectivity test systems has shown that SSRs can be rather different in nature. A variety of SSRs present in a selectivity set is


a
Selectivity ratio
> 50 for cat K
Selectivity for cat S and K
< 10-fold
Potency for cathepsin K [pKi or pIC50]
Potency for cathepsin S [pKi or pIC50]
b
S-selectives
Non-selectives
K-selectives
Potency for cathepsin K [pKi or pIC50]
1/1000
1/50
1/10
10
50
1000
Selectivity for K over S, K/S [SR]
Set 1
Set 2
c
Non-selective ligand
cathepsin K
cathepsin S
1.3 nM
1.6 nM
1
2
Molecule 1 binds cathepsin K with a 1.2-fold higher potency compared to cathepsin S → Selectivity ratio (SR) 1.2 for K/S
Selective ligand
cathepsin K
cathepsin S
1 nM
52 nM
1
2
Molecule 2 binds cathepsin K with a 52-fold higher potency compared to cathepsin S → Selectivity ratio (SR) 52 for K/S

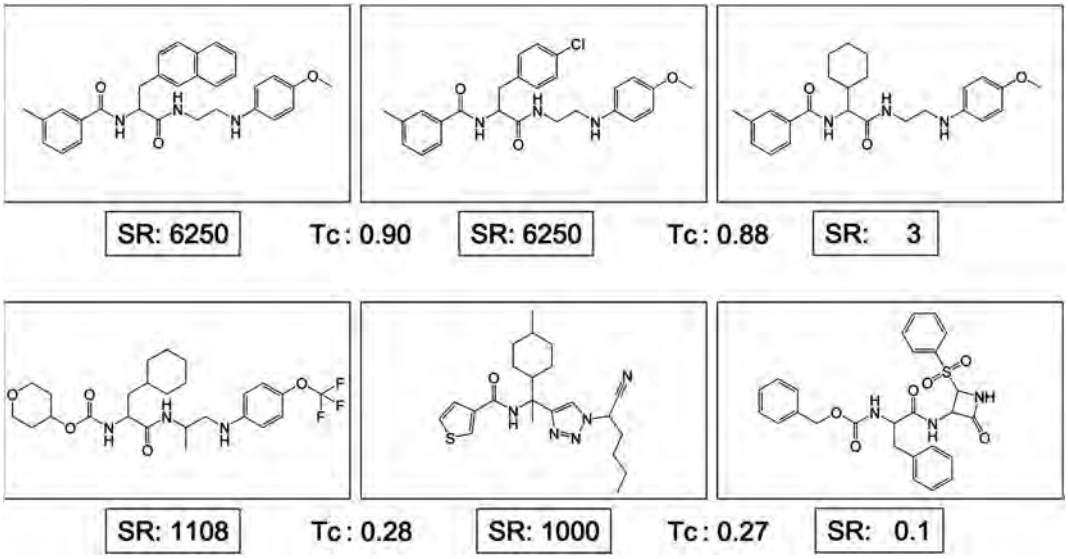


Fig. 3. Structure–selectivity relationships. Different types of structure–selectivity relationships are illustrated that are present in a set of cathepsin enzyme inhibitors. Structurally similar as well as diverse molecules display different selectivity patterns. SR (S/K) is reported. As a measure of structural similarity, "Tc" reports pair-wise MACCS Tanimoto coefficient values. Inhibitors are taken from a "cat S/K 7-8" set.

shown in Fig. 3. Structurally similar as well as structurally diverse compounds are found to be selective for the same target and ligands with distinct selectivity are related to each other by different degrees of structural similarity. The compound series in Fig. 3 shows that structural analogs might display the same selectivity (top left) or, in contrast, significant differences in their selectivity profile (top right). Moreover, diverse structures are found to exhibit comparable selectivity (bottom left) or different selectivity (bottom right). All of these compounds belong to a cathepsin S over K (S/K) selectivity set and show comparable potency for cathepsin S (12–48 μM). Thus, for compounds in the first row, minor structural changes do not affect the potency for cathepsin S but only the selectivity for S over K. These findings suggest that there are no simple structural rules that govern selectivity differences.

Fig. 2. Classification of selective and nonselective compound. (**a**) Inhibitors of cathepsin (cat) K and S are shown. Potency values against cat S are plotted against the corresponding values for cat K on the pK_i and pIC_{50} scale. Diagonals delineate the 50-fold potency intervals. Compounds selective for cat K over S are colored *black,* and nonselective compounds with less than tenfold difference in potency are shown in *gray*. *Horizontal lines* delineate two potency ranges for compounds included in different selectivity set. (**b**) Potency values of compounds for cat K are plotted against the corresponding SR values for cat K over S. Only molecules with potency values within an order of magnitude for cat K (here pK_i and pIC_{50} from 7 to 8 or 8 to 9) are retained in a selectivity set comprising selective (*black dots*) and nonselective (*gray dots*) molecules. (**c**) An exemplary molecule from each set is shown and the calculation of SR illustrated.

2.4. Selectivity Searching

The compound benchmark systems reported here have been designed to enable the development and evaluation of computational methods to assess compound selectivity. In benchmark investigations, both selectivity sets have been used to evaluate computational methods including 2D molecular fingerprints, mapping algorithms, or support vector machines for their potential to distinguish between selective, active but nonselective, and inactive database compounds [11, 15–17]. These methods are indeed found to significantly enrich selective over nonselective and inactive compounds, suggesting that existing methodologies can be well adapted for selectivity searching and SSR analysis.

3. Selectivity Profile Mining

The sets described in Subsection 2.2 focus on selectivity differences between different target pairs. For the extension of selectivity analysis to multiple targets, a methodological framework is required that enables the treatment of complex and less well-defined selectivity relationships. Therefore, a data mining approach, Formal Concept Analysis (FCA) has been adapted that is capable of formulating and exploring complex selectivity relationships and profiles.

Formal Concept Analysis is a data mining and visualization technique [18] originating from computer science that we have been recently utilizing for the systematic mining of structure–activity relationships [19] and selectivity profiles [20]. This section describes MolFCA and how it is used for the definition and mining of complex selectivity profiles.

3.1. Formal Concept Analysis

FCA operates on binary relationships of the general form "is a" or "is not a," e.g., "Aspirin is a cyclooxygenase inhibitor" or "Aspirin is not a cyclooxygenase inhibitor." All such relationships are reported in a *formal context*. Figure 4a shows a schematic representation of a formal context. It represents a table where rows correspond to objects (e.g., molecules) and columns to attributes (e.g., bioactivity annotations). A cross is placed in a cell if an object has a certain attribute. *Formal concepts* are defined as sets of objects that share a specific subset of attributes. FCA systematically extracts concepts from a given formal context. Concepts are not independent of each other: objects that share attributes "A" and "B" also belong to the concept that is defined by attribute "A" alone. These relationships between concepts are visualized in so called *concept lattices*, as illustrated in Fig. 4b. Each node in a concept lattice corresponds to a concept. Attributes are written above nodes, and objects are written below nodes. In order to find the attributes and objects associated with any concept, the edges

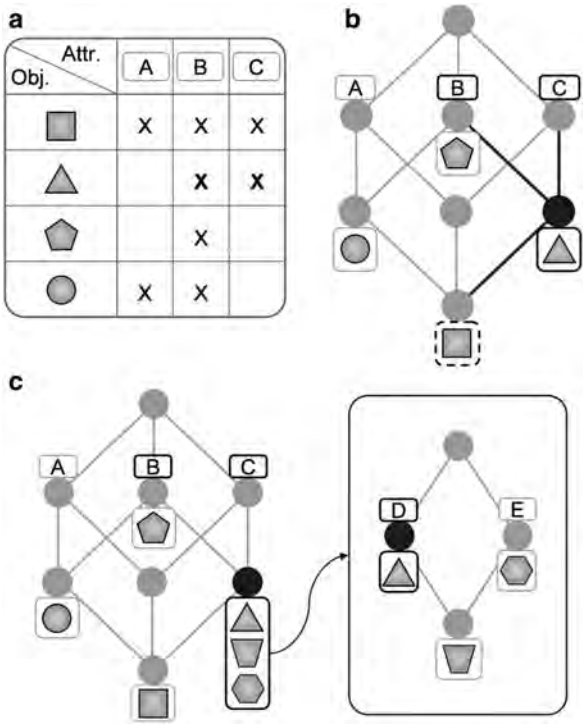


Fig. 4. Formal Concept Analysis. (**a**) A formal context is shown that consists of four objects and three attributes. An "x" is placed in a cell if an object has the respective attribute. (**b**) The concept lattice for the context is shown. Each node represents one concept. The concept that is defined by attributes "B" and "C" is selected. The triangle is directly associated with the node because it has attributes "B" and "C," but not "A." By tracing the edges of the lattice, all subjects belonging to the concept can be identified. The square also shares attributes "B" and "C," but is associated with the bottom node because it also has attribute "A." (**c**) Two scales are shown that account for different subsets of attributes. The combination of scales forms a query. In this example, three objects are selected on the first scale and projected onto the next scale, which shows their distribution for attributes "D" and "E".

of the lattice are traced. Attributes are identified by following the edges toward the top node and objects by following edges toward the bottom node. Thus, concept lattices show the distribution of objects across different attributes.

3.2. Scales and Queries

For formal contexts with a large number of attributes (i.e., many columns in the table), concept lattices can become difficult to navigate because the number of concepts grows exponentially with the number of attributes. Therefore, *scales* are designed that focus on defined subsets of attributes, e.g., ligand activity annotations for closely related targets or target families [19]. Thus, each scale reports the distribution of objects across attributes that are grouped based on domain knowledge. For example, in Fragment Formal Concept Analysis (FragFCA) [19], different scales are used

to assess the distribution of fragment combinations across ligands with closely related targets. Therefore, scales are defined for individual target families. Other scales are designed to assess ligand potency for each individual target. In order to build queries that span different sets of attributes, scales are combined. Objects corresponding to a concept of interest can be selected on one scale and projected onto the next scale, as shown in Fig. 4c. For interactive assembly of a query, scales serve two purposes. First, they report the distribution of objects across different groups of attributes, e.g., different target families or potency ranges. Second, each scale allows the selection of relevant objects and scale-based queries focusing on a subset of objects that are of specific interest.

3.3. Molecular Formal Concept Analysis

In addition to structure–activity relationship analysis, FCA has also been adapted for the systematic mining of selectivity profiles in biologically annotated databases [20]. Scale design and scale-based queries are well suited for the definition of complex selectivity profiles, given the pair-wise nature of compound selectivity assessment. As discussed above, ligand selectivity is also defined here on the basis of potency ratios against two targets at a time. Compounds that are selective for one of several targets are distinguished from nonselective ligands using potency ratio thresholds (e.g., $\geq$50-fold for selective and $\leq$10-fold for nonselective compounds). Thus, three selectivity compound sets can be distinguished for a target pair A and B: compounds selective for A over B, compounds selective for B over A, and nonselective compounds for A and B, corresponding to the design of compound selectivity sets, as discussed above. In biologically annotated databases, compound potency information is often available only for a subset of targets, i.e., not all compounds have been tested against all targets. This is often the case for public domain databases that are assembled from literature sources. Therefore, compounds with no available potency information are distinguished from inactive compounds. Therefore, MolFCA utilizes selectivity scales in order to distinguish different compound selectivity criteria and account for missing data points, due to a lack of potency information. A prototypic selectivity scale is shown in Fig. 5a. In MolFCA, attributes correspond to targets and objects to molecules. Selectivity scales show the distribution of compounds among the three selectivity sets. Compounds that show comparable potency are associated with both targets and written below the bottom node, whereas selective ligands are associated with a single target. Molecules that lack potency information or compounds that cannot be classified as selective or nonselective using the two thresholds do not have any attributes and are written directly below the top node. Such selectivity scale is generated for each pair of targets in a database. In addition to selectivity, MolFCA also incorporates activity information. Therefore, potency scales have been designed

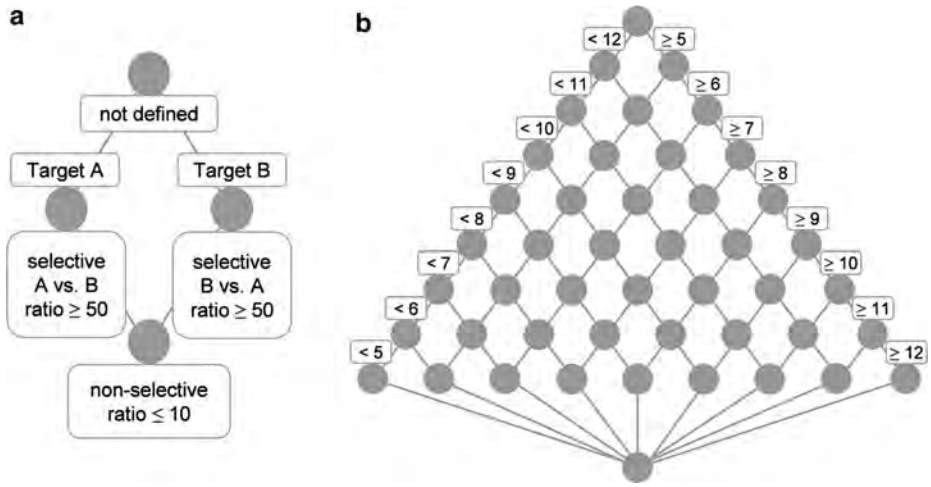


Fig. 5. Scales for Molecular Formal Concept Analysis. (**a**) A prototypic selectivity scale is shown. Attributes are target annotations, and molecules are assigned to different nodes based on their selectivity against the two targets. Selectivity scales are generated for each pair of targets. (**b**) A prototypic potency scale is shown that reports the distribution of molecules over different potency intervals. A potency scale is generated for each individual target.

that assess the distribution of compounds across different potency ranges on a logarithmic scale. Potency scales are complementary to selectivity scales because two compounds can differ in their activity but still show the same selectivity against two targets. A prototypic potency scale is shown in Fig. 5b.

3.4. Selectivity Profile Mining

The combination of selectivity scales and potency scales enables the definition of complex selectivity and potency profiles. Molecules selected on one selectivity scale can be assessed on selectivity scales that focus on a distinct target pair. As discussed above, each scale shows the distribution of compounds among the three selectivity categories or potency bins. Subsets of molecules are selected and projected onto further scales. Given a template compound, the associated nodes (i.e., concepts) are selected on each scale, leading to the identification of compounds sharing the same selectivity profile. Figure 6a shows an exemplary query for histone deacetylase inhibitors using vorinostat [21] as a template. As shown in Fig. 6b, the identified compounds can be structurally rather diverse, because MolFCA utilizes only potency and selectivity annotations as molecular descriptors and is hence not structurally biased in any way.

3.5. Query Modification and De Novo Queries

The interactive query assembly using scales allows modification of queries in a defined manner. A query can be modified by adding or removing scales or by selecting different nodes on individual scales. Figure 6c depicts compounds that have been identified through query modification and deviate from the vorinostat profile. Comparison of structurally similar compounds that have different

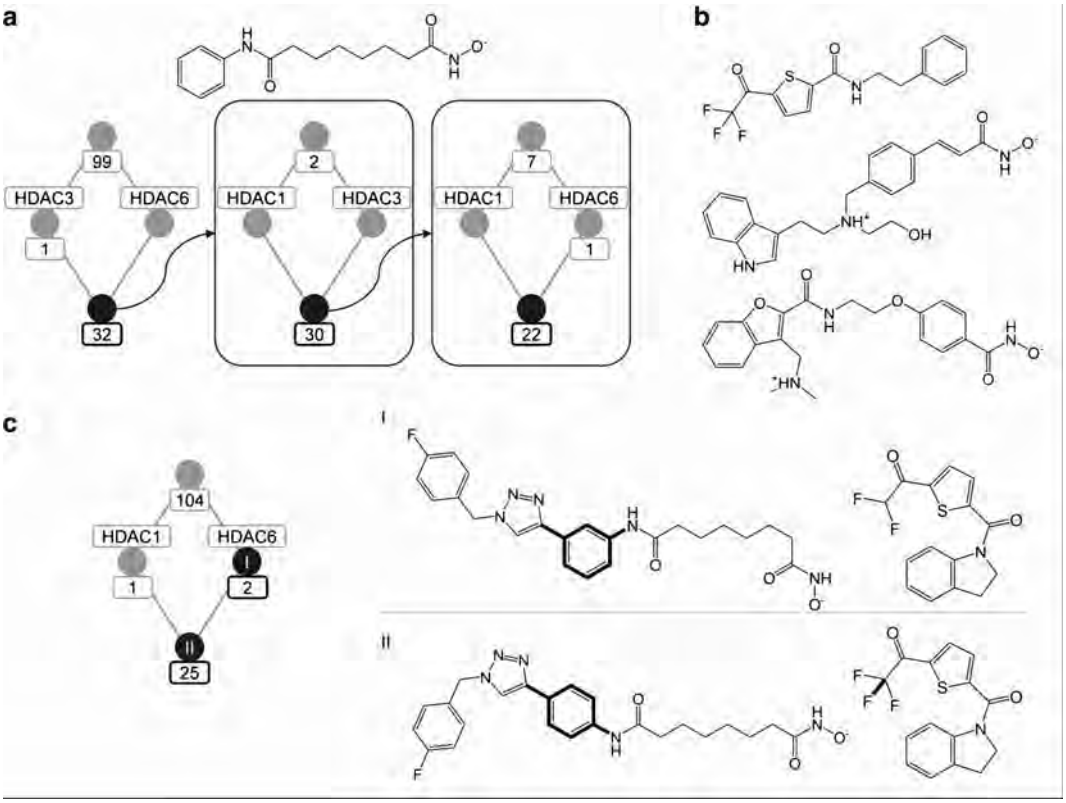


Fig. 6. Selectivity profile queries. (**a**) A histone deacetylase (HDAC 1, 3, and 6) inhibitor selectivity query based on vorinostat (*top*) is shown. Three scales are combined, and each time the concept that contains vorinostat is selected. On the first scale, 32 compounds are identified that are further reduced to 30 compounds on the second and to 22 compounds on the third scale. (**b**) Structurally diverse compounds are shown that have been identified using the vorinostat selectivity profile query. (**c**) Two structurally similar pairs of compounds with different selectivity are shown identified by selecting different concepts on the scale depicted on the *left*. Compounds selected from node I deviate from the vorinostat profile and are structurally similar to compounds from node II, which correspond to vorinostat selectivity.

selectivity profiles makes it possible to assess SSRs in complex selectivity profiles. In Fig. 6c, structural determinants of selectivity for HDAC6 over HDAC1 are highlighted in structurally very similar compounds. The utilization of selectivity scales that rely on the same criteria also permits the definition of queries in the absence of template compounds. Because selectivity is assessed using well-defined threshold values, robust de novo selectivity queries can be assembled that represent a selectivity profile of interest without the need to obtain a representative template compound.

3.6. Summary

Binary compound selectivity systems have been designed to enable the evaluation of computational methods for selectivity analysis in the context of chemical biology applications. Furthermore, MolFCA extends pair-wise selectivity assessment by systematic definition and mining of selectivity profiles. Compounds are described by their potency and selectivity against a panel of

targets. In order to define a complex query, selectivity and potency scales are combined, and compounds selected on one scale are projected onto the next scales. Therefore, MolFCA provides a standardized methodological framework for the definition and exploration of compound selectivity profiles in biologically annotated databases.

References

1. Jacoby, E. (2006) Chemogenomics: drug discovery's panacea? *Mol. Biosyst.* **2**, 218–220.
2. Spring, D. R. (2005) Chemical genetics to chemical genomics: small molecules offer big insights. *Chem. Soc. Rev.* **34**, 472–482.
3. Bajorath, J. (2008) Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery. *Expert. Opin. Drug Discov.* **3**, 1371–1376.
4. Tan, D. S. (2005) Diversity-oriented synthesis: exploring the intersections between chemistry and biology. *Nat. Chem. Biol* **1**, 74–84.
5. Stockwell, B. R. (2004) Exploring biology with small organic molecules. *Nature* **432**, 846–854.
6. Bredel, M., and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5,** 262–275.
7. Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. (2006) Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805–815.
8. Yildirim, M. A., Goh, K., Cusick, M. E., Barabási, A., and Vidal, M. (2007) Drug-target network. *Nat. Biotechnol.* **25**, 1119–1126.
9. Cheng, Y., and Prusoff, W. H. (1973) Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **22**, 3099–3108.
10. Stumpfe, D., Ahmed, H. E. A., Vogt, I., and Bajorath, J. (2007) Methods for computer-aided chemical biology. Part 1: design of a benchmark system for the evaluation of compound selectivity. *Chem. Biol. Drug Des.* **70**, 182–194.
11. Stumpfe, D., Geppert, H., and Bajorath, J. (2008) Methods for computer-aided chemical biology. Part 3: analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem. Biol. Drug Des.* **71**, 518–528.
12. Johnson, M. A., and Maggiora, G. M. (1990) Concepts and applications of molecular similarity. Wiley, New York.
13. Kubinyi, H. (1998) Similarity and dissimilarity – a medicinal chemist's view. *Perspect. Drug Discov. Des.* **11**, 225–252.
14. Peltason, L., and Bajorath, J. (2007) Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.* **14**, 489–497.
15. Vogt, I., Stumpfe, D., Ahmed, H. E. A., and Bajorath, J. (2007) Methods for computer-aided chemical biology. Part 2: evaluation of compound selectivity using 2D molecular fingerprints. *Chem. Biol. Drug Des.* **70**, 195–205.
16. Wassermann, A. M., Geppert, H., and Bajorath, J. (2009) Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.* **49**, 582–592.
17. Vogt, I., Ahmed, H. E. A., Auer, J., and Bajorath, J. (2008) Exploring structure-selectivity relationships of biogenic amine GPCR antagonists using similarity searching and dynamic compound mapping. *Mol. Divers.* **12**, 25–40.
18. Priss, U. (2006) Formal concept analysis in information science. *Annu. Rev. Inform. Sci. Technol.* **40**, 521–543.
19. Lounkine, E., Auer, J., and Bajorath, J. (2008) Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds. *J. Med. Chem.* **51**, 5342–5348.
20. Lounkine, E., Stumpfe, D., and Bajorath, J. (2009) Molecular Formal Concept Analysis for compound selectivity profiling in biologically annotated databases. *J. Chem. Inf. Model.* **49**, 1359–1368.
21. Duvic, M., Talpur, R., Ni, X., Zhang, C., Hazarika, P., Kelly, C., Chiao, J. H., Reilly, J. F., Ricker, J. L., Richon, V. M., and Frankel, S. R. (2007) Phase 2 trial of oral vorinostat (suberoylanilide hydroxamic acid, SAHA) for refractory cutaneous T-cell lymphoma (CTCL). *Blood* **109**, 31–39.

Chapter 21

Application of Support Vector Machine-Based Ranking Strategies to Search for Target-Selective Compounds

Anne Mai Wassermann, Hanna Geppert, and Jürgen Bajorath

Abstract

Support vector machine (SVM)-based selectivity searching has recently been introduced to identify compounds in virtual screening libraries that are not only active for a target protein, but also selective for this target over a closely related member of the same protein family. In simulated virtual screening calculations, SVM-based strategies termed preference ranking and one-versus-all ranking were successfully applied to rank a database and enrich high-ranking positions with selective compounds while removing nonselective molecules from high ranks. In contrast to the original SVM approach developed for binary classification, these strategies enable learning from more than two classes, considering that distinguishing between selective, promiscuously active, and inactive compounds gives rise to a three-class prediction problem. In this chapter, we describe the extension of the one-versus-all strategy to four training classes. Furthermore, we present an adaptation of the preference ranking strategy that leads to higher recall of selective compounds than previously investigated approaches and is applicable in situations where the removal of nonselective compounds from high-ranking positions is not required.

Key words: Target-selectivity, Selectivity searching, Machine-learning, Support vector machines, Kernels, Multi-class SVM, SVM preference ranking

1. Introduction

With the advent of chemical biology and chemogenomics, small molecules have become increasingly important tools for studying and elucidating biological functions [1, 2]. To systematically explore the functions of related protein targets, it is of high relevance to identify small molecules with different selectivity patterns against individual members of protein families [3]. To support experimental efforts, several methods including conventional similarity searching and advanced machine-learning approaches have been evaluated for their potential to enrich small database selection sets with target-selective compounds

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_21, © Springer Science+Business Media, LLC 2011

and deselect active but nonselective molecules to ensure high *purity* of the database selection set [4, 5]. The most promising strategies reported so far are based on *support vector machines* (SVM) [6–8], an algorithm originally developed to solve binary classification problems. In chemoinformatics, a popular application is the classification of active vs. inactive compounds [9, 10]. To these ends, SVM has proven to be a powerful approach because, instead of only trying to minimize the classification error on the training data, the algorithm also employs so-called structural risk minimization methods in order to avoid over-fitting effects and enhance generalization performance. In a typical chemoinformatics SVM application, training compounds belonging to two different classes are projected into chemical feature space, and the SVM subsequently derives a hyperplane in this space to separate the two classes. Furthermore, the use of kernel functions in SVM learning enables the classifier to derive a more complex (nonlinear) decision boundary and generalize to cases in which the two classes are not linearly separable. Test compounds are classified based on which side of the decision boundary they fall. Although originally developed for making yes/no decisions, SVMs can also be utilized to generate a ranking of database compounds [11, 12]. However, selectivity searching provides additional challenges for conventional SVM analysis. Instead of two training classes, three classes containing selective, nonselective (but active), and inactive molecules are required for training, corresponding to a three-class ranking problem. Therefore, SVM multi-class strategies termed *preference* and *one-versus-all ranking* have recently been introduced for the identification of target-selective molecules [5].

In this chapter, we present an extension of the one-versus-all SVM strategy to treat four instead of three classes. In order to predict compounds that are selective for a target *A* over *B*, SVM learning has been based on training compounds belonging to three categories: target *A*-selective, nonselective (i.e., molecules with comparable potency for target *A* and *B*), and (assumed to be) inactive, as mentioned above. Here we show that the inclusion of a fourth category consisting of target *B*-selective compounds can further improve search performance by increasing the ratio of target *A*-selective over nonselective molecules. This ratio (*purity*) can be utilized to assign a high probability of target-selectivity to a compound. However, one might also aim at identifying as many target-selective compounds as possible, regardless of the number of nonselective compounds that are also detected. For this purpose, the preference ranking strategy can also be adapted to primarily focus on recovery of selective compounds, rather than purity.

We test this alternative preference ranking strategy on six different ligand sets that contain target-selective as well as nonselective compounds and compare it to original preference ranking

and standard SVM ranking that uses selective compounds and decoys for training. Furthermore, an application example of the one-versus-all strategy using four training classes is presented for cathepsin (Cat) inhibitors.

2. Methods

In the following, we describe and compare methodological details of standard SVM-based ranking, preference ranking, and the one-versus-all strategy and compare virtual screening results. In all calculations, MACCS structural keys [13] were used as compound descriptors.

2.1. Support Vector Machines

SVM are a supervised machine-learning technique [7, 8]. A computational model is built based on a training set to associate class labels with particular feature patterns. As a binary molecular classification approach, SVM learning makes use of training data $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$ with $\mathbf{x}_i \in \mathbf{R}^d$ being the fingerprint representation of a molecule i (a bit vector) and $y_i \in \{-1,+1\}$ being its class label (positive or negative) to deduce a hyperplane H that best separates positive from negative training examples. The hyperplane H is defined by a normal vector $\mathbf{w}$ (with Euclidean norm $||\mathbf{w}||$) and a scalar b (called bias) so that

$$H: \langle \mathbf{x}, \mathbf{w} \rangle + b = 0 \tag{1}$$

where $\langle \cdot, \cdot \rangle$ defines a scalar product.

Given a linearly separable training set, an infinite number of hyperplanes exist to correctly classify the data. The particular hyperplane chosen by an SVM is the one maximizing the distance (called *margin*) from the nearest training examples, which is a basic requirement for good generalization performance. Without loss of generality, the constraints to be met by the training data for correct classification can be expressed as the following set of inequalities:

$$y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 \quad \forall i \tag{2}$$

The compounds for which equality holds in Eq. 2 are closest to the hyperplane H and are termed *support vectors*. The distance from H to the support vectors from the positive and the negative training class is $1/||\mathbf{w}||$ meaning that maximizing $1/||\mathbf{w}||$ or minimizing $||\mathbf{w}||$, given the conditions specified in Eq. 2, returns the maximum margin hyperplane. If the training examples are not linearly separable, the optimization problem has no solution. To solve this problem, the strict constraints in Eq. 2 are softened by insertion of positive slack variables ξ_i, allowing for training

examples to fall within or on the incorrect side of the margin. The value of the slack variable ξ_i correlates with the distance from the incorrectly positioned training compound i to the edge of the corresponding class-specific margin. The following minimization problem arises by introducing a new parameter C as a cost factor that penalizes training errors and is adjustable to find a compromise between training accuracy and the size of the margin:

minimize:

$$V(\mathbf{w}, \xi) = \|\mathbf{w}\|^2 + C \sum_i \xi_i \tag{3}$$

subject to:

$$y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i \quad \text{with} \quad \xi_i \geq 0 \quad \forall i \tag{4}$$

Optimization problems under constraints can be solved by the introduction of Lagrange multipliers [6], which yields a convex quadratic programming problem amenable to standard methods:

maximize:

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{5}$$

subject to:

$$\sum_i \alpha_i y_i = 0 \quad \text{with} \quad 0 \leq \alpha_i \leq C \quad \forall i \tag{6}$$

The solution vector $\boldsymbol{\alpha}$ determines the normal vector $\mathbf{w}$ of H as a linear combination $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$. Since only support vectors (i.e., those vectors falling on the edge, within, or on the incorrect side of the margin) obtain factors α_i greater than zero, the position of the hyperplane exclusively depends on these critical vectors. For all other vectors, α_i is equal to zero. If a solution to the optimization problem formulated in Eqs. 5 and 6 has been found, and $\mathbf{w}$ and b have been determined, a test molecule $\mathbf{x}$ is classified on the basis of the decision function

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \right) \tag{7}$$

This means that compounds with $f(\mathrm{x}) = 1$ are assigned to the positive class and those with $f(\mathrm{x}) = -1$ to the negative class. Geometrically, the sign reflects on which side of the hyperplane a test molecule falls.

In many applications, a planar surface might not be capable of separating the data correctly. This problem can be eliminated by a transformation of the data into a high-dimensional space $\mathcal{H}$ where a linear separation of the data might become feasible. If one assumes

that the projection is accomplished using a mapping Φ: $\mathbf{R}^d \rightarrow \mathcal{H}$, then the optimization problem given in Eqs. 5 and 6 requires the calculation of dot products in $\mathcal{H}$, i.e., functions of the form $\langle \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2) \rangle$. However, by finding suitable kernel functions $K(\mathbf{x}_1, \mathbf{x}_2)$ that correspond to scalar products in $\mathcal{H}$, i.e., $K(\mathbf{x}_1, \mathbf{x}_2) = \langle \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2) \rangle$, the explicit calculation of the embedding function Φ can be circumvented, which is referred to as the *kernel-trick* [8]. The embedding function Φ does not have to be known because the decision function $f(\mathbf{x})$ can also be extended to the use of kernels:

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \right) \tag{8}$$

To adapt the approach to virtual compound screening, the classification method should be transformed into a ranking function. An intuitive way to do so is to define the rank of a molecule i (with fingerprint representation $\mathbf{x}_i$) according to the signed distance of its embedding $\Phi(\mathbf{x})$ to the hyperplane determined in $\mathcal{H}$. For the compounds that fall in the half-space populated by class -1, the distance is converted to its negative value such that the compounds are ranked from the most distant compound located on the side of the positive training class to the most distant compound on the side of the negative training class. This ranking methodology is equivalent to removing the sgn function in Eq. 8 and sorting the molecules in descending order of

$$g(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) \tag{9}$$

Hence, the simplest SVM-based approach to search for selective compounds is training with selective molecules as the positive class and inactive molecules as the negative class and ranking a test database according to Eq. 9.

2.2. Original Preference Ranking

Preference ranking [14] was originally developed to optimize the retrieval quality of search engines. For this purpose, the algorithm derives a ranking function based on pairwise preference constraints (e.g., "document d_i should rank higher than document d_j"). It is straightforward to show that preference ranking leads to an optimization problem that is equivalent to SVM classification [14]. In order to adapt this concept for selectivity searching, two preference constraints must be formulated: (1) selective molecules (class S) should rank higher than inactive and nonselective molecules (class I and N, respectively) and (2) inactive molecules should obtain a higher rank than nonselective molecules, which is expressed in the binary relationship

$$R = (\mathrm{S} \times \mathrm{N}) \cup (\mathrm{S} \times \mathrm{I}) \cup (\mathrm{I} \times \mathrm{N}) \tag{10}$$

The underlying idea is to learn a linear ranking function g depending on a weight vector $\mathbf{w}$ that sorts the test molecules according to the constraints defined by R:

$$\forall(i,j) \in R: \quad \langle \mathbf{x}_i, \mathbf{w} \rangle > \langle \mathbf{x}_j, \mathbf{w} \rangle \tag{11}$$

This corresponds to deriving a normal vector $\mathbf{w}$ for a set of parallel hyperplanes dividing the space into ordered layers of selective, inactive, and nonselective molecules. Test compounds are then sorted by their projection onto $\mathbf{w}$ (Fig. 1).

If a perfect linear separation is not possible, violations of preference constraints are permitted by the introduction of slack variables $\xi_{i,j}$ and a cost factor C for margin regularization, analogously to the binary classification SVM described in Subsection 2.1. These modifications result in a convex optimization problem that can be solved similarly to the one of the SVM classification problem (note the direct correspondence to Eqs. 3 and 4):

minimize:

$$V(\mathbf{w}, \xi) = \|\mathbf{w}\|^2 + C \sum_{i,j} \xi_{i,j} \tag{12}$$

subject to:

$$\langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{w} \rangle \geq 1 - \xi_{i,j} \quad \text{with} \quad \xi_{i,j} \geq 0 \quad \forall(i,j) \in R \tag{13}$$

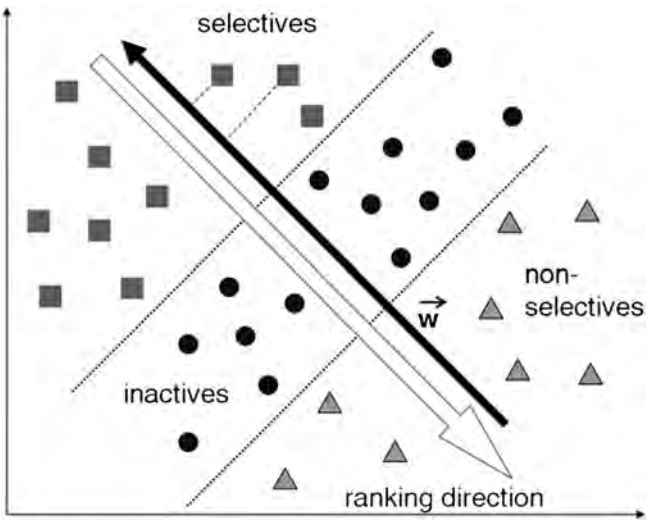


Fig. 1. *Preference ranking*. Preference ranking uses the three training classes simultaneously to learn a weighting vector **w** that is perpendicular to a set of parallel hyperplanes best separating selective from inactive and inactive from nonselective training molecules. Test compounds are ordered by their projection onto **w** (as exemplarily indicated for two selective compounds by *dashed lines*).

Factors α_l are determined by the Lagrangian reformulation of this problem and define the weight vector $\mathbf{w}$ as a linear combination of the training vectors $\mathbf{x}_l$: $\mathbf{w} = \sum_l \alpha_l \mathbf{x}_l$.

The algorithm is also extendable to nonlinear decision functions, i.e., to the insertion of kernels, so that the ranking function $\mathcal{J}$ is given by

$$\mathcal{J}(\mathbf{x}) = \sum_l \alpha_l K(\mathbf{x}_l, \mathbf{x}) \tag{14}$$

2.3. Original One-Versus-All

The one-versus-all approach is a commonly used strategy to make binary classification methods applicable to multi-class problems. In the context of selectivity searching, three individual "one-versus-rest" SVM-based ranking functions $\mathcal{J}_S, \mathcal{J}_N, \mathcal{J}_I$ (similar to Eq. 9) are derived for the three training classes selective (S), nonselective (N), and inactive (I) by training with classes S, N, or I vs. the union N ∪ I, S ∪ I, or S ∪ N, respectively, as visualized in Fig. 2. The test set is sorted three times in descending order of the values given by each of the three ranking functions such that a test molecule j obtains three individual ranking positions $\text{rank}^S(j)$, $\text{rank}^N(j)$, and $\text{rank}^I(j)$. Then, the class assignment for j is made by

$$\underset{X \in \{S,I,N\}}{\arg\min}\ \text{rank}^X(j) \tag{15}$$

under the assumption that a molecule is most likely to belong to the class for which it ranks best (*see* **Note 1**). To generate a final ranking for molecules classified as selective, they are again sorted by their values of $\mathcal{J}_S$.

In systematic test calculations on 18 compound data sets, both the one-versus-all strategy and preference ranking produced promising results [5]. In comparison to standard SVM utilizing selective molecules and decoys for learning, these more advanced approaches produced comparable recovery rates of selective test compounds but database selection sets of higher purity (*see* **Note 2**). These strategies were evaluated in combination with different 2D fingerprints and kernel functions, and it was shown that combining preference ranking with the Gaussian kernel [15] yielded highest purity, whereas one-versus-all ranking was the overall more robust method (*see* **Note 3**) [5].

2.4. Adaptation of Preference Ranking to Increase Recovery of Selective Compounds

As nonselective compounds are often structurally very similar to selective molecules, the use of nonselective compounds also as positive training examples might further increase the recall of selective compounds (but reduce the purity of selection sets). However, simply combining selective and nonselective compounds for training is inappropriate if one aims to distinguish between selective and nonselective molecules. Rather, one could use preference ranking with the priority order changed to

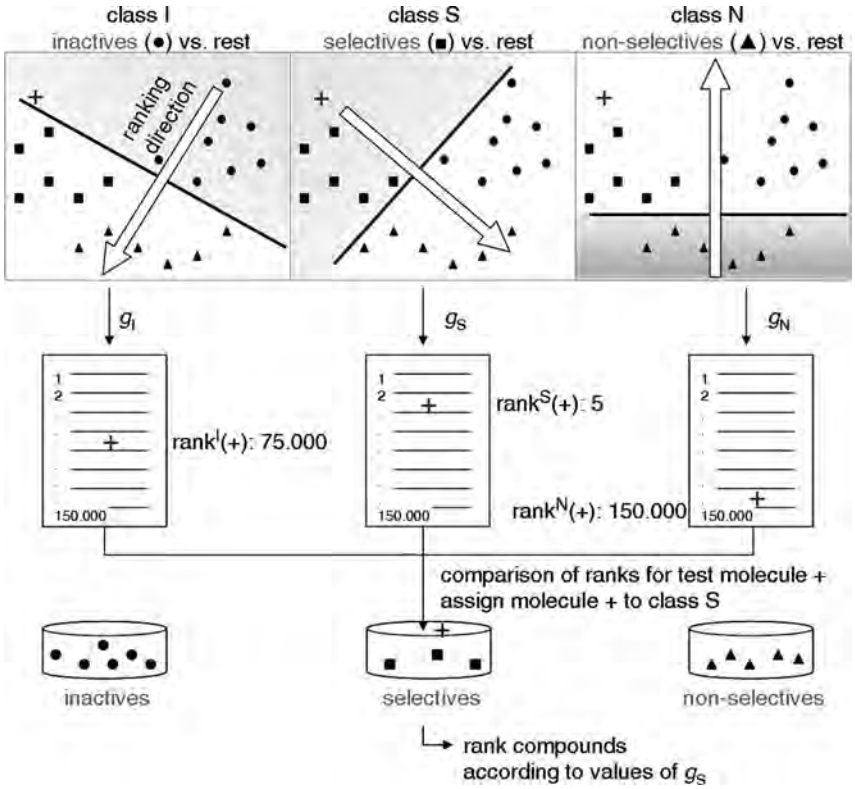


Fig. 2. *One-versus-all*. Here the multi-class classification problem is divided into three binary classification problems such that SMV learns to separate each class from the union of the two remaining classes. Test compounds are subjected to each binary classification task and three separate rankings are produced. Then, for each test compound (+), its three positions in the individual rankings are compared and a compound is assigned to the class for which it ranks highest. In order to generate a final ranking, test molecules assigned to the selective class (class S) are sorted by their initial scores for S.

selective – nonselective – inactive, i.e., selective molecules should obtain a higher rank than nonselective molecules and nonselective molecules a higher rank than decoys. This would ensure that selective molecules with high structural similarity to nonselective compounds would retain a comparably high-rank position (*see* also **Note 4**).

2.5. Extension of One-Versus-All to Four Training Classes

In order to enrich a database selection set with molecules that are selective for a target *A* over *B* and remove compounds that are nonselective for *A* and *B*, a fourth training class is used for one-versus-all ranking consisting of compounds that are selective for target *B* over *A*. This modification is also motivated by the often observed structural similarity of nonselective and selective compounds. The introduction of the fourth training class might be expected to preferentially deprioritize nonselective molecules that rank higher for the "selective for *B*-versus-rest" classifier than for the "selective for *A*-versus-rest" classifier.

2.6. Application Examples

2.6.1. Evaluation of Alternative Preference Ranking

In order to test the alternative version of preference ranking in simulated virtual screening trials, we have carried out a comparative study using standard SVM ranking and the original version of preference ranking as a reference. We have used six previously described selectivity sets [4] containing between 12 and 31 compounds that are selective for a particular target over a closely related one as well as between 10 and 35 compounds that are active but nonselective for these two targets (*see* also **Note 5**). These sets are reported in Table 1. In light of our previous observation that preference ranking with its original priority order performed best in combination with the Gaussian kernel, we also used this kernel function. In the learning phase, half of the selective and nonselective molecules of each data set were used as training molecules and the remaining compounds were added to the MDL Drug Data Report (MDDR) [16] that was used as the background database for selectivity searching in this case (*see* **Note 6**). Furthermore, 100 compounds randomly taken from the MDDR served as putative decoys (inactive training examples). For each selectivity set, ten independent search trials with randomly chosen training sets were carried out. As performance measure for comparing the alternative preference ranking strategy with the two reference methods, recovery rates were calculated for selection sets consisting of either 100 or 1,000 database compounds and averaged over the ten independent trials.

Table 1 summarizes the results of this study. As can be seen, alternative preference ranking failed to improve recovery rates for small database selection sets of 100 compounds. In comparison to standard SVM ranking, in part significant decreases in recovery rates (up to Δ14%) were observed for four selectivity sets, whereas only two sets showed a slight increase in the number of recovered selective molecules. Perhaps unexpectedly, original preference ranking also performed overall better than alternative preference ranking, despite its comparably poor performance for the trypsin/thrombin set (*see* Table 1). However, for selection sets of 1,000 compounds, alternative preference ranking yielded best recovery rates for four selectivity sets. The highest increase was observed for the set Cat K/B of Δ18% in comparison to standard SVM ranking and Δ23% compared to original preference ranking. Hence, these results indicated that preference ranking with the priority ordering *selective – nonselective – inactive* had the potential to increase the number of identified selective compounds in database selection sets of medium size.

2.6.2. Evaluation of the Extended One-Versus-All Strategy

To test the extension of one-versus-all ranking to four training classes, a set of cathepsin K and S inhibitors was assembled. This set consisted of 41 nonselective inhibitors, 45 inhibitors selective for cathepsin K, and 39 inhibitors selective for cathepsin S. For the modified one-versus-all strategy, four one-versus-rest classifiers

Table 1
Comparison of alternative versions of preference ranking with standard SVM ranking in simulated virtual screening trials

Selectivity set	Number of selectives	Number of nonselectives	Method	Recovery 100	Recovery 1,000
CA IX/I	14	11	Standard	88.6	97.1
			Original	88.6	95.7
			Alternative	81.4	95.7
Cat K/B	12	10	Standard	38.3	48.3
			Original	38.3	43.3
			Alternative	41.7	66.7
Cat K/L	31	10	Standard	60.0	69.4
			Original	59.4	69.4
			Alternative	53.1	80.6
Cat S/L	14	13	Standard	52.9	67.1
			Original	58.6	68.6
			Alternative	48.6	60.0
MMP 8/1	12	16	Standard	53.3	76.7
			Original	57.5	65.0
			Alternative	55.0	83.3
Trypsin/ thrombin	13	35	Standard	42.9	62.9
			Original	17.1	31.4
			Alternative	28.6	71.4

"Selectivity set" reports the names of the six target pairs studied in our calculations: *CA* carbonic anhydrase, *Cat* cathepsin, *MMP* matrix metalloprotease. The designation "CA IX/I" means that the set consists of molecules that are selective for CA IX over CA I and also nonselective molecules for these targets. For each data set, the number of selective and nonselective compounds is given. Under method, "standard" means standard SVM ranking, "original" for preference ranking with the priority order *selective – inactive – nonselective*, and "alternative" for preference ranking with the priority order *selective – nonselective – inactive*. Recovery rates are reported for database selection sets of 100 ("Recovery 100") and 1,000 compounds ("Recovery 1,000").

were built, i.e., "selective for Cat K-versus-rest", "selective for Cat S-versus-rest", "nonselective-versus-rest", "inactive-versus-rest". Predictions for test molecules were then analyzed from two points of view: (1) Preferential identification of inhibitors selective for cathepsin K. Therefore, test molecules that ranked best for the "selective for Cat K-versus-rest" classifier were sorted according to their initial score for this classifier, and purity and recovery rates were determined for the resulting compound list. (2) Preferential

identification of inhibitors selective for cathepsin S. Hence, purity and recovery rates were calculated for those molecules that ranked best for the "selective for Cat S-versus-rest" classifier. As reference methods for the extended one-versus-all strategy, one-versus-all calculations with three training classes (i.e., selective for the target of interest, nonselective, inactive) and standard SVM ranking with two training classes (selective, inactive) were carried out. In analogy to the calculations described above, the Gaussian kernel function was used, and ten independent search trials were carried out for all three strategies. For each trial, ten training compounds were randomly selected from each selective and nonselective subset, and the remaining compounds served as potential database hits. In addition, 100 putative decoys were taken from the MDDR as inactive training examples. Search results were averaged over all trials and are reported in Fig. 3 for database selection sets of 1,000 compounds.

In Fig. 3, the histograms on the right show that comparable recovery rates were achieved for all three methods. However, with respect to purity, the two different one-versus-all approaches clearly outperformed standard SVM ranking. Moreover, when searching for selective cathepsin K inhibitors, adding a fourth training class further increased purity from ~90% to 95%. Thus, a number of nonselective molecules that were classified as selective for cathepsin K when only using three training classes were no longer present in the final compound list because they were classified as selective for cathepsin S.

Nonselective inhibitors in our data set were structurally more similar to inhibitors selective for cathepsin S than cathepsin K. Accordingly, when searching for selective cathepsin S inhibitors,

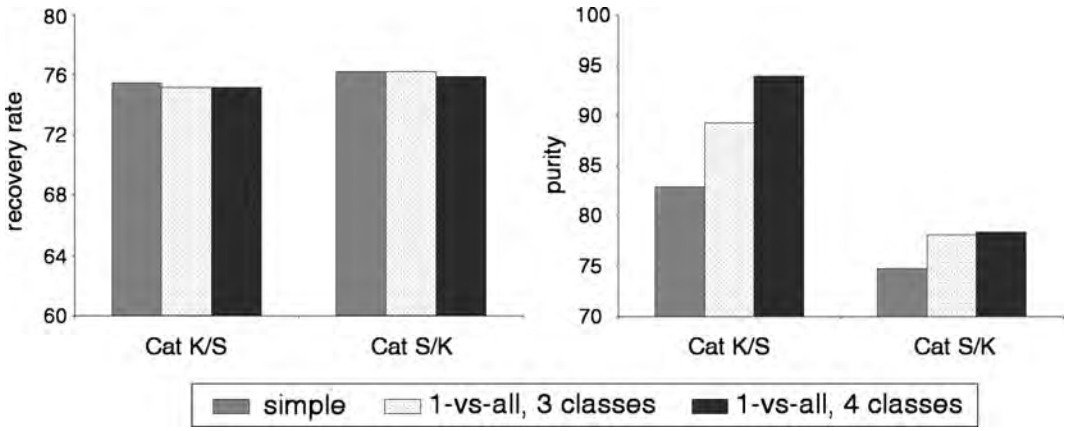


Fig. 3. *Comparison of recovery rate and purity for different SVM approaches.* Standard SVM ranking is compared to the one-versus-all approach that utilizes either three or four training classes. Recovery rates and purities are reported for inhibitors of cathepsins K and S in database selection sets of 1,000 compounds. Results are averaged over ten independent search trials with different randomly assembled training sets.

the inclusion of a training class with compounds selective for cathepsin K could not further improve purity by filtering out "misclassified" nonselective inhibitors. This example shows that we cannot expect that the extension of the one-versus-all strategy to a fourth training class consistently improves SVM search performance in all cases. Rather, the approach is thought to complement existing strategies for selectivity searching.

3. Notes

1. A direct comparison of scores (i.e., $g_S(j)$, $g_N(j)$, $g_I(j)$) was considered inappropriate due to the different space representations utilized in the different SVM classifiers. Replacing scores with ranks is often applied to make different search calculations comparable.
2. *Recovery rates* report the number of identified selective molecules relative to the total number of selective compounds available in the screening database. *Purity* is defined as the number of recovered selective molecules divided by the total number of recovered active (i.e., selective and nonselective) molecules.
3. Because preference ranking places nonselective test molecules toward the end of the final compound list, selective test molecules that are highly similar to nonselective training molecules might easily be "misclassified" and deprioritized, leading to reduced recovery rates. Indeed, a dramatic decrease in recovery rates was observed when the preference ranking strategy was used in combination with low-complexity 2D fingerprints (i.e., TGD [17] and MACCS) and the linear kernel [18] operating in comparably low-dimensional feature spaces. However, when combined with the Gaussian kernel, preference ranking yielded high recovery rates and excellent purities for all investigated fingerprints (GpiDAPH3 [17], MACCS, Molprint2D [19], TGD). Regardless of the kernel function used, the one-versus-all strategy always produced recovery rates that were comparable to those achieved by standard SVM ranking, but consistently improved purity.
4. Pooling selective and nonselective training compounds can also lead to further improved recall of selective compounds. However, test calculations have shown that the success of this strategy highly depends on the ratio of selective versus nonselective reference molecules. There should be at least as many selective as nonselective molecules in the training set to avoid that the decision function is predominantly determined by

features inherent to nonselective molecules. However, reducing a large nonselective subset to a size that is equal to or smaller than the selective subset means that the available structural information is not fully utilized. This problem can be circumvented by using preference ranking that assigns a higher priority to selective molecules, irrespective of the number of selective and nonselective molecules in the training set.

5. Selectivity was defined as a potency ratio of at least 50 (ratio of K_i or IC_{50} values for a pair of targets) whereas a potency ratio between 0.1 and 10 was applied as a criterion for nonselectivity.
6. The MDDR stores structural and activity data for biologically relevant compounds assembled from the scientific or patent literature. Known inhibitors for the ten target proteins considered in our study and inhibitors for closely related targets as well as molecules with identical 2D graphs were removed from the database, resulting in an MDDR version of 152,336 compounds.

References

1. Bredel, M. and Jacoby, E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5**, 262–275.
2. Stockwell, B. R. (2004) Exploring biology with small organic molecules. *Nature* **432**, 846–854.
3. Bajorath, J. (2008) Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **12**, 352–358.
4. Stumpfe, D., Geppert, H., and Bajorath, J. (2008) Methods for computer-aided chemical biology, part 3: analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem. Biol. Drug. Des.* **71**, 518–528.
5. Wassermann, A. M., Geppert, H., and Bajorath, J. (2009) Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.* **49**, 582–592.
6. Burges, C. J. C. (1998) A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121–167.
7. Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
8. Boser, B. E., Guyon, I. M., and Vapnik, V. A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, 1992; ACM: New York, 1992; pp 144–152.
9. Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **26**, 5–14.
10. Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., and Lemmen, C. (2003) Active learning with support vector machines in the drug discovery process. *J. Chem Inf. Comput. Sci.* **43**, 667–673.
11. Jorissen, R. N. and Gilson, M. K. (2005) Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **45**, 549–561.
12. Geppert, H., Horváth, T., Gärtner, T., Wrobel, S., and Bajorath, J. (2008) Support-vector-machine-based ranking significantly improves the effectiveness of similarity searching using 2D fingerprints and multiple reference compounds. *J. Chem. Inf. Model.* **48**, 742–746.
13. MACCS Structural Keys; Symyx Software: San Ramon, CA, 2005.
14. Joachims, T. Optimizing search engines using clickthrough data. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, 2002; ACM: New York, 2002; pp 133–142.

15. Powell, M. J. D. Radial basis functions for multivariable interpolation: a review. In Mason, J. C., and Cox, M. G. (eds). *Algorithms for Approximation*; Clarendon Press, Oxford: 1987; pp 143–167.
16. *MDL Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2005.
17. *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
18. Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. (2005) Graph kernels for chemical informatics. *Neural Netw.* **18**, 1093–1110.
19. Bender, A., Mussa, H. Y., Glen, R. C., and Reiling, S. (2004) Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **44**, 170–178.

Chapter 22

What Do We Know?: Simple Statistical Techniques that Help

Anthony Nicholls

Abstract

An understanding of simple statistical techniques is invaluable in science and in life. Despite this, and despite the sophistication of many concerning the methods and algorithms of molecular modeling, statistical analysis is usually rare and often uncompelling. I present here some basic approaches that have proved useful in my own work, along with examples drawn from the field. In particular, the statistics of evaluations of virtual screening are carefully considered.

Key words: Statistics, Central Limit Theorem, Variance, Standard deviation, Confidence limits, *p*-Values, Propagation of error, Error bars, *logit* transform, Virtual screening, ROC curves, AUC, Enrichment, Correlation, Student's *t*-test, ANOVA

1. Introduction

About 3 years ago I was shocked to discover I didn't know anything. This self-realization came as a bit of a surprise. At that time I was working on the problem of how to generate shapes that might fit onto the surface of a protein and had come to a point where I realized to do this "rigorously" I would need to apply some statistics. I wasn't sure what I really meant by "rigorous," I only knew what it did not mean, i.e., making up parameter values that seems "good enough". I was tired of reading and reviewing papers that took that approach; indeed I was tired of the general lack of anything fundamental in most of modeling. It would be great if we could solve problems by applying physics to drug discovery, but those wise in the ways of the field know our problems are typically too messy for this. Eventually and inevitably, physics based approaches will work, but for now the only real alternative is statistics, and how hard can that be?

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_22, © Springer Science+Business Media, LLC 2011

It turned out to be a lot harder than I thought. Since my initial revelation, I've attempted to remedy my ignorance and would find my previous cluelessness hard to believe if it were not for the day in, day out evidence of the general statistical naivety of those who practice molecular modeling. This concerns me because I would like our field to be effective, one that contributes as much as possible to the most important industry on earth – the discovery of these amazing small molecules with their potential for dramatic effects on health and well-being. Furthermore, if you do not understand statistics your publications may not mean anything, you are unlikely to add to the body of knowledge that defines the field, you probably won't know if you are making progress and, in general, are missing one of the best ways to rationally comprehend the world around you.

A full accounting of all the things I have learnt since my decision to stop being ignorant would not fit into this chapter. Rather, I will present a more practical agenda of some mostly simple tests that I have found useful and applied in my own work. These range from the very simplest concepts of error bounds, to the more complicated ideas of significance tests and validation metrics. This survey is by no means meant to be comprehensive – my own ignorance is still profound! As such, I will end with suggested readings for those interested in delving either more deeply or broadly.

2. Statistics

2.1. Types of Statistics and Why They Matter

One of the first and most confusing things an explorer in statistics finds is that there are different kinds of statistics. The biggest divide is between what is called "frequentist" statistics and Bayesian statistics. Nearly all of the statistics we learn is of the frequentist variety; typically, we "discover" Bayesian statistics later in life and then wonder why we did not know this stuff. The frequentist approach is to consider what would happen if you could repeat an experiment over and over again and look at the set of results you generate. I remember this was the first Physics experiment we were made to do in my undergraduate days. The Bayesian approach is built around the concept of prior probabilities and how observations, even singular ones, alter our perceptions of these probabilities. It especially appeals to physicists because it is deceptively simple and decidedly general, whereas the myriad of statistical tests and approaches from the frequentist approach often seem messy and arbitrary. The philosophical differences between the two approaches are fascinating and I recommend a little known article by T. J. Loredo [1] on this debate and also on the Bayesian approach in general. We won't say more about the

Bayesian approach here, although a good treatise on its application to modeling is long overdue. Rather, I am going to concentrate on that "myriad of statistical tests." Why? Because those tests are very useful and also appeal to the physicist in me, if for a different reason. If you look at most mathematical physics of the first half of the twentieth century it was dominated by clever mathematics. How else could you get a result when there were no computers? Similarly, the classical statistics outlined here is the consequence of practical people wanting rules of thumb that worked. As a result an impressive patchwork of understanding and clever methods evolved. These usually enable "back-of-the-envelope" calculations. Bayesian analysis only gained respect and utility in the second half of that century because computers became available. Although there are some clever methods to work analytically with the Bayes approach, e.g., using Gaussian priors, most of the work requires computational resources unavailable to the founders of modern (frequentist) statistics. From a practical perspective, both have their place.

Within the world of frequentist statistics there is another divide, that between "parametric statistics" and "nonparametric statistics". Parametric statistics simply means that we are making assumptions about an underlying *probability distribution*, i.e., the distribution of results we would get if we could redo that experiment as many times as we liked. The function that describes this distribution is called a *p*robability *d*ensity *f*unction, or *pdf* – an unfortunate acronym given the document format from Adobe, but statisticians were there first. A *pdf* is a continuous function, often a Gaussian, and the estimation of the "parameters" of that function is usually the name of the game. Now, not all distributions are going to be continuous, e.g., flipping a coin only gives heads or tails, nothing in between. In these cases, we have to use discrete distributions, e.g., the probability of heads and the probability of tails. However, we are often interested in averages, e.g., what is the distribution of the fraction of heads seen after many flips and as such *pdfs* often represent the limiting form derived from discrete distributions after large numbers of observations.

Nonparametric methods make less assumptions because they do not care what the distribution looks like. These methods are particularly useful for things like ranked lists. An example would be the Kendall Tau, a method of assessing the quality of a ranking of two types of object (e.g., active/inactive) that I find very useful. Again, a good treatment of the application of nonparametric statistics is overdue in our field. However, one of the objects we shall concern ourselves with is the ROC (Receiver Operator Characteristic) curve, and its associated Area Under the Curve (AUC), a nonparametric measure of retrieval rate originally formulated for radar analysis, i.e. we shall be looking at the parametric statistics of a nonparametric measure!

The cornerstone of parametric statistics, what I shall call "classical" statistics, is the Central Limit Theorem (CLT). This remarkable theorem asserts that the average of enough measurements has a *pdf* that asymptotically becomes a Gaussian. Most of this article is an elaboration of what you can do with this one result, for instance, obtaining error bounds, something usually missing from modeling reports. I consider where the Gaussian approach falls down, in particular when the sample size is small and when the range of values of the average is bounded and the actual average is near one of these bounds. I will then discuss the propagation of error, i.e., how multiple sources of error contribute to an overall bound. Seemingly a simple concept, the application of this idea to determine intrinsic variance of methods is under-appreciated. I will show how this can be applied to the study of virtual screening evaluation. This naturally leads to a consideration of the errors in measuring the difference of two quantities, in some ways the main use of classical statistics, i.e., is method A or drug A better than method B or drug B. This, in turn, provides a backdrop to the Student's *t*-test, i.e., whether a set of results contains two rather than one subpopulation. I'll describe but not go into details of how this can be generalized to the famous ANOVA method of Fisher that dominates the landscape of multivariate analysis prevalent in the social sciences. This leaves a lot out that is important, such as chi-squared tests, *F* functions, experimental design, contingency tables, sampling methods, cross-validation, robust statistics and, perhaps most significantly, regression. But this is a chapter, not a book.

2.2. Classical Statistics and the Central Limit Theorem

The most basic and perhaps useful of all theories in statistics concerns the average value of something. The Central Limit Theorem (CLT) states that no matter what the distribution of a quantity with a random component, if you average enough independent measurements of the same system, the average property will be distributed like a Gaussian. These conditions, "independent", "identical" are so common in statistical theory that they go by the abbreviation "i.i.d" – "independently and identically distributed". (We might indeed wonder if we are ever so lucky as to have "i.i.d" samples in computational chemistry!) The center of a CLT Gaussian is the true mean and its width depends on something called the *variance*. Variance is just expected average of the square of the difference of a property from its average value.

When you stop and think about it this is a strange and wonderful result. It doesn't matter what the underlying distribution of the quantity might be, for instance, it might only have two possible values and not be a continuous distribution. What matters is that the average follows a Gaussian, with all a Gaussian's well-understood properties. For being such a profound theory it is not actually difficult to prove and the only real requirement is that the

variance of the distribution function be finite. If we follow typical conventions, the mean is represented by the letter μ and the variance by σ^2. We write the variance as the square of another quantity because σ is often referred to as the standard error or standard deviation and is just the square root of the variance.

$$\text{best estimate of mean} = \mu = <x> = \frac{1}{N}\sum_{i=1}^{N} x_i$$
$$\text{best estimate of variance} = \sigma^2 = \text{var} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{1}$$

You might ask why the average of the squared deviations from the mean involves $N-1$, rather than N and, in fact, this is something that often confused me. First of all, it should not matter very much. William Press, he of "Numerical Recipes" [2] fame, once wrote, "If using N rather than $N-1$ makes a difference to your results then you are already misusing statistics". The reason $N-1$ is used rather than N is because the mean we use here is our *estimate* of the mean from the same data – if we somehow knew the exact mean then we would more properly use N and not $N-1$. It turns out that if we have to use μ rather than the true mean we tend to underestimate the true variance, i.e., as N gets bigger we approach the true variance from below. Using $N-1$ makes the calculated variance an "unbiased estimator" of the true variance, i.e., as N gets big we approach the true value with equal likelihood from above or below. Another way to think of this is that there are really only $N-1$ observations in the variance because any of the measurements can be recalculated from the other observations and the mean, i.e., there are really $N-1$ *degrees of freedom*. Finally, the way I remember it is from what happens when $N=1$. If we somehow know the true mean we can still calculate a variance because we divide by "1". If we do not then the variance is $(0/0)$, i.e., 0 on top because the mean is the same as the single value, and 0 below because $N-1$ is 0. As such, we do not mistakenly believe the variance is zero; rather it is undefined, which is appropriate for the variance of a single-observation.

Another way of writing the variance is as follows:

$$\begin{aligned}\sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i^2 - 2x_i\mu + \mu^2)\\ &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i^2 - \mu^2)\\ &\approx <x^2> - <x>^2\end{aligned} \tag{2}$$

Where the angled brackets just mean "average." This will turn out to be a particularly useful expression when we come to deal

with the variances of probabilities. One final item concerning variances will come in handy – suppose we change variables from x to some new convolution of x represented by a function f, i.e., $y = f(x)$. What will we be able to say about the variance of y, if we know the variance of x? A good approximation to this, on average, turns out to be just:

$$\sigma_y^2 \approx \left(\frac{df}{dx}\right)_{x=\mu}^2 \sigma_x^2 \tag{3}$$

With estimates of both the mean and the variance, the CLT tells us just what the probability distribution, i.e., *pdf*, of the average of N observations looks like.

$$pdf\ (x) = \sqrt{\frac{N}{2\pi\sigma^2}}\, e^{-N(x-\mu)^2/2\sigma^2} \tag{4}$$

Here, μ is the measured mean and the prefactor, i.e., the number that multiplies the exponential, is there so that if we integrate the *pdf* from negative to positive infinity we get 1.0, i.e.,

$$\int_{-\infty}^{\infty} \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(x-\mu)^2/2\sigma^2} dx = 1 \tag{5}$$

Figure 1 shows you what a Gaussian looks like where we have set N and σ to 1.0 and μ to 0.0. The way we interpret this distribution is that the most likely value of the average is the mean, i.e., the center of the Gaussian, but if we were to make

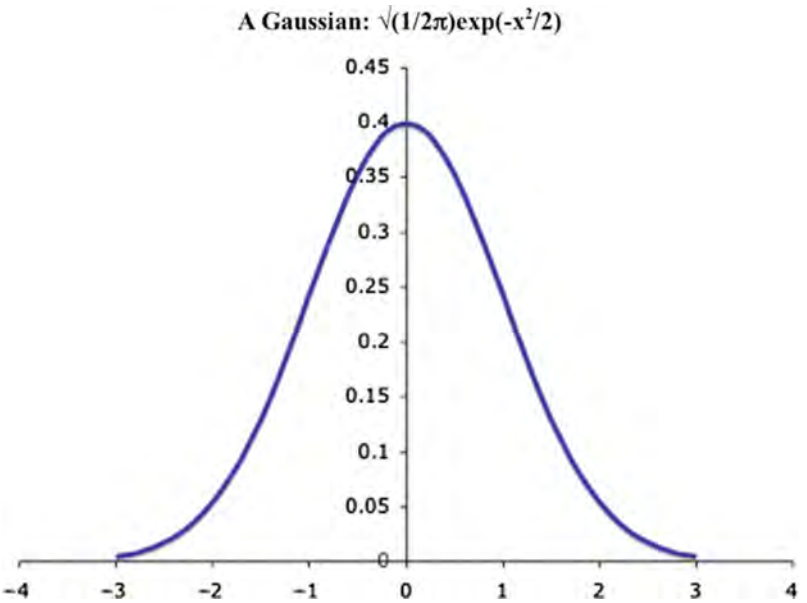


Fig. 1. Example of a Gaussian form.

another N measurements we would likely get a different mean and it would be distributed according to this *pdf*. Remember, though, if we make a single additional observation, observation $N + 1$, this is not the *pdf* for this new observation. It is the *pdf* for the average of another N measurements. This said, many observations themselves have a Gaussian *pdf*, i.e., are drawn from such a distribution. For example, human height forms a Gaussian distribution. Single observations Gaussian *pdf*s are common for characteristics that have multiple influences, i.e., are the average themselves of multiple other parameters, effectively arriving at a CLT result in a single measurement. It is very handy when this is the case, but not essential for what follows.

The most important characteristic of this *pdf* is the interplay between the *variance* and N, the number of observations. As the variance goes up, i.e., the uncertainty in any one measurement, the *pdf* gets broader. As the number of observations goes up the *pdf* gets narrower, i.e., more observations mean more certainty. In the Gaussian, the distance from the mean (the "error") enters as a squared property, as does the standard deviation; however, N enters as a single (reciprocal) power. This is the root of why error goes down as the square root of the number of observations. De Moivre first deduced this in 1731 from the properties of the binomial distribution and as such the essence of the CLT is often credited to him. It is one of the most approachable and often least appreciated of mathematical results [3]. Lack of such knowledge can be costly. Coinage in medieval England, *circa* 1150 AD, was subject to the "Trial of the Pyx" – essentially 100 coins were chosen at random from a mint and compared to a standard weight [4]. As minting coins was known to be imperfect an allowance was made such that the average coin could be heavier or lighter by a certain amount. However, because the $\sqrt{N}$ rule was not known it was assumed the net weight of the 100 coins should be allowed to vary by this same fraction! This meant unscrupulous mints could routinely shave off gold from some coins, secure in the knowledge that the average would still pass the Trial.

If we want quantitative estimates of the error, rather than just "better" or "worse," we have to consider the properties of a Gaussian. The probability of any particular range of values is simply the area under the Gaussian in that range. What range of values around the average would we consider "likely"? The standard procedure is to state that we want 95% of all experiments to fall within such a range. To get the range equivalent to this percentage we need to know how the area under a Gaussian grows as a range expands from the center. We call this a *cumulative distribution function*, often written as "*cdf*." The *cdf* for a Gaussian is called the *erf* function, precisely because it is the *er*ror *f*unction. Actually *erf* is defined slightly differently from a general *cdf* in that it is the area under the Gaussian from zero (center) to

some positive value, rather than from negative infinity to some finite value:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt \tag{6}$$

Since the Gaussian is symmetric the number we want is the value of x such that $erf(x) = 0.95$, i.e., 95% of the distribution greater than zero lies within the range $[0, x]$

$$x = erf^{-1}(0.95) \tag{7}$$

The correct value of x is about 1.386. If we compare the result from the CLT we see that we have to rescale the x in the exponential by $\sqrt{(N/(2\sigma^2))}$ to make it look like the *erf* function, i.e.,

$$\begin{aligned} x &= 1.386\sqrt{2\sigma^2/N} \\ &= 1.96\sqrt{\sigma^2/N} \\ &\approx 2\sqrt{\sigma^2/N} \end{aligned} \tag{8}$$

This then is the classic result that gives us a simple estimate of the 95% confidence limit for a reported value. It is so useful that the square root term is simply known as the "standard error of the mean" (SEM), i.e.,

$$\begin{aligned} \text{SEM} &= \sqrt{\sigma^2/N} \\ &= \sigma/\sqrt{N} \end{aligned} \tag{9}$$

A few of questions spring to mind – isn't all this well-known and obvious? Perhaps, however, the frequency of error bars for values reported in computational presentations is very low. At a recent ACS meeting I visited every poster on computational chemistry, over 40 in all, and only one had values with confidence limits displayed. This is changing as more become aware of the importance of demonstrating some statistical knowledge, especially when trying to publish.

A second question might be whether we want both sides of the Gaussian. For instance, suppose we are interested in the average being less than some threshold with 95% confidence. For instance, in the Trial of the Pyx the King was perfectly happy with the 100 coins weighing more than they were supposed to, only coming in too light had serious consequences. In this case we do what is called, reasonably enough, a "one-sided *t*-test". We are only interested in the threshold that has 95% above it, rather than the two thresholds either side of the mean that contain 95% of the weight. In this case, half of the weight is already accounted for above the mean and we instead require:

$$x = erf^{-1}(0.90) \tag{10}$$

This leads to an x of 1.645, rather than 1.96, i.e., as expected you go out less far to get to the point where 5% of the values are below the "one-sided" threshold. Mostly we will be interested in the two-sided test, i.e., region around the mean. We won't be concerned with whether deviations from the mean are on the high side or the low side; we just want to know if the deviation is "significant." An example of a two-sided test is whether a variant of a method has an impact, i.e., does the performance lie within expected variation for that difference – we don't ask whether the impact is for better or worse. A one-sided test would be to explicitly ask whether method A is now better than method B.

A final question might be, "Why 95% instead of 80% or 99%?" The answer to this is historical. R. A. Fisher is the father of most modern statistics. He began his work while at the Rothamsted Experimental Station in Hertfordshire, England in 1919. Rothamsted is an agricultural research institute and Fisher was studying crop yields. Essentially, our addiction to 95% comes directly from Fisher deciding that fields with fertilizers added ought, on average, to have more cabbages than all but one in 20 unfertilized fields. From such earthy beginnings are we stuck with 95% as a measure of significance!

2.3. Deviations from the CLT Limit

2.3.1. Finite Sample Size

Many statistical theories are called "asymptotic." What this means is that they hold if the number of samples is big enough. The Central Limit Theorem is no exception. For the systems we are likely to be interested in we will be nowhere near the asymptotic limits of any statistical theory, so how do we deal with finite samples? Unlike the limiting behavior this will depend on the single-observation *pdf*. However, suppose this *pdf* is itself a Gaussian, as it often is for physical observables. The first to solve this problem in 1908 was William Sealy Gosset [5], better known to the world of statistics as "Student." He had to publish under this name because Guinness, his employer, did not want the world to know they were using statistics! "Student" of course, invented the "Student's t-test," which uses the Student's t-distribution, the function we want. Gosset showed if we have $n + 1$ observations then the function we want looks like:

$$\begin{aligned} f_n(t) &= c_n(1 + t^2/n)^{-(n+1)/2} \\ c_{n_\text{even}} &= \frac{(n-1)(n-3)..(3)(1)}{2\sqrt{n}(n-2)(n-4)..(4)(2)} \\ c_{n_\text{odd}} &= \frac{(n-1)(n-3)..(4)(2)}{\pi\sqrt{n}(n-2)(n-4)..(3)(1)} \end{aligned} \tag{11}$$

As you can see in Fig. 2 the larger n is the more the function looks like a Gaussian, as the CLT would require, but we can now

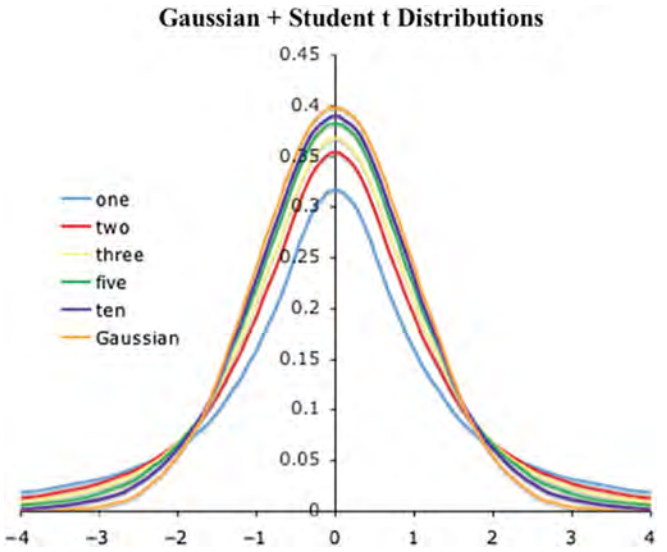


Fig. 2. Examples of the Student's *t*-distribution.

Table 1
Example Student's *t*-values for different degrees of freedom corresponding to 95% confidence limits (two tail tests)

Degrees of freedom	Critical *t*-value
1	12.706
2	4.303
5	2.571
10	2.228
25	2.060
50	2.009
100	1.984
∞	1.960

deal with smaller samples. Practically, what does this mean for confidence limits? Essentially, it means the "1.96" we use to get (two-sided) 95% confidence limits needs to be adjusted. There are standard tables for this and I've added a few example values in Table 1. As you see, "1.96" starts larger, e.g., 2.23 when $n = 10$ and slowly, asymptotically, approaches 1.96. Remember that the value we use is one less than the number of observations. Again, there is no distribution function if there is only one observation, i.e., there is no "0" Student function.

2.3.2. Extrapolation from the Mean

There's another way in which the main result of the CLT can be misleading – the Gaussian form is accurate close to the true average but not necessarily far away. In some sense this is obvious – a Gaussian *pdf* has probabilities from negative to positive infinity and most measurements do not go that far. When you measure a weight, for instance, the probability is zero that the weight is negative. Similarly, if we are measuring a probability we can only have a result between zero and one. Since we use the tails of the Gaussian to place our confidence limits this can lead to some stupid looking error bars. What do we do? There are really two possibilities, and I will discuss them here because we are going to use one of them later. The simple one, the one we aren't going to use, is simulation. The reason we aren't going to use this is because this article is all about how to estimate without simulation. However, we do all have computing power beyond the imaginings of the founders of classical of statistics, so sometimes it is not a bad idea to use it. An example of how to do this is to take the data that you started with and sample from it "with replacement". What this means is that we select a data point and then "put it back," i.e., it is possible we will pick it again as the second data point. Keep doing this until you have a dataset the size of the original and calculate the new average. Do this a lot of times to create a spread of values (the average of which should be close to that of your original set) and then find the range that includes 95% of all such trials. Crude, but effective. Now for the second way. Remember if we change variables, i.e., make up some function of the primary variable x, then we can still estimate the variance of the new variable. Suppose our original variable is bounded between two limits. We can always scale these limits to 0 and 1, so let's do so. Now consider the function:

$$y = l(x) = \log\left(\frac{x}{1-x}\right) \tag{12}$$

This is called the *logit* function and it maps the range $[0, 1]$ to $[-\infty, \infty]$. This means that the function l has the same range as a Gaussian. If the variance of x is σ^2 then, from our previous formula, the variance of l is:

$$\begin{aligned} \mathrm{var}_l &\approx \left(\frac{\mathrm{d}l}{\mathrm{d}x}\right)_{x=\mu}^2 \mathrm{var}_x \\ &\approx \left(\frac{1}{\mu\,(1-\mu)}\right)^2 \mathrm{var}_x \end{aligned} \tag{13}$$

So now all we do is to calculate the error bars for the function l, instead of x, using this new variance, safe in our knowledge there are no boundaries to run into. We then take these confidence

limits of l, and work out the values of x corresponding to each. This is easy because the inverse *logit* function is just:

$$l^{-1}(y) = \frac{1}{1 + e^{-y}} \tag{14}$$

Sounds complicated, but trust me it is a lot better than having error bars that look silly. Appendix 1 gives more details about the process and I will give a worked example.

2.4. Compound Variances

There are really two main reasons to quantifying error. The first is if we are calculating or measuring something that will be used as a part of something else. For instance, if we were trying to estimate the speed of light that value is going to be used in other quantities of interest. Or, closer to home, suppose we wanted to use a prediction method in conjunction with other methods and we want to know what the error of the combined method would be. The second use of errors is in comparing quantities. This is perhaps more usual in the biomedical field, e.g., we want to know if drug A is better than drug B, or in the computational world, whether method A gives better results than method B. These two problems are actually very related. In the first case, we are asking how errors add, i.e., when quantities are combined, in the second case we are asking how they "subtract," i.e., when looking at the differences of quantities. The golden rule for errors is simply that the net error is the square root of the sum of the errors squared, or, more simply, "variances add." For instance, if we are combining methods A and B into a then the expected standard deviation for C is as follows:

$$\sigma_{A+B} = \sqrt{\left(\sigma_A^2 + \sigma_B^2\right)} \tag{15}$$

This central result follows naturally from the fact that independent probabilities multiply. *See* Appendix 2 for more details. The same basic rule applies for both difference and sums – remember that when we change x to a function of x the square of the derivative of the new function scales the variance, i.e., whether we are adding or subtracting the sign for combination is always positive. Understanding how errors compound really lies at the heart of the practical use of the CLT and classical statistics.

2.4.1. An Example Using the AUC

We might not be combining anything. There might simply be independent sources of error in a measurement that add as the square root of the sum of their squares. An excellent example of this can be found in the consideration of an AUC, i.e., the Area Under the Curve, of a ROC curve. *See* Fig. 3 as an example of an AUC of 0.75. ROC curves are very popular in a number of fields where recall of a small set of "actives" is required from a large background of "inactives." If "events" are ordered by some value, the ROC curve is the line representing the fraction of actives

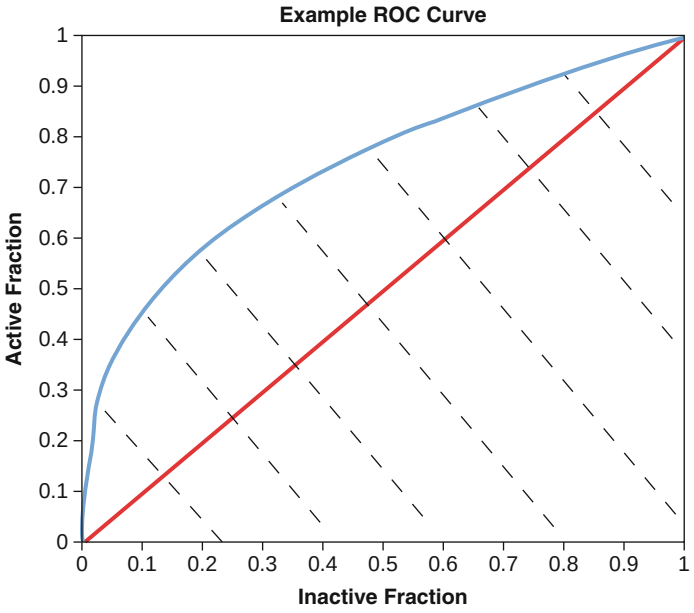


Fig. 3. An example of a ROC curve. *Upper line*: AUC = 0.75. *Lower line*: AUC = 0.5. *Dashed area* is the Upper AUC, i.e., area under the curve.

(y-value) at any point in the list with the fraction of inactives (x-value) at that point. The area under this curve is the AUC and represents the probability that a randomly selected active is higher in the list than a randomly selected inactive. The nice thing about the AUC is that it does not depend on the number of actives or inactives or their ratio, unlike some traditional evaluation measures. That makes it a property of the method, not the experiment. However, the error in the AUC naturally depends on these numbers. The intriguing thing to consider is that it has to depend on both numbers, i.e., the finiteness of the number of actives and of the number of inactives are both sources of error.

If we have the complete list (of actives and inactives) we actually have a couple of ways to calculate the AUC error bounds. Since the AUC is the average of a probability of a given active being higher than all the inactives, we can just calculate the variance of this average. This might lead us to think, that the SEM of the AUC is:

$$\begin{aligned} \mathrm{SEM\ (AUC)} &= \sqrt{\frac{1}{N_\mathrm{A}} \sum_{i=1}^{N_\mathrm{A}} (p_\mathrm{i} - < p >)^2 / (N_\mathrm{A} - 1)} \\ &= \sqrt{\frac{1}{N_\mathrm{A}} \sum_{i=1}^{N_\mathrm{A}} (p_\mathrm{i} - \mathrm{AUC})^2 / (N_\mathrm{A} - 1)} \end{aligned} \qquad (16)$$

where N_A is the number if actives, p_i is the probability active i is higher in the list than a random inactive and in the second line we

have replaced the average probability, $<p>$, with, by definition, the AUC. To see why this is wrong we have to remember that the "p" we calculate for each active has an error associated with it because the number of inactive is finite, i.e., we have not included the error from the inactives. DeLong [6] showed the correct formula is:

$$\text{SEM (AUC)} = \sqrt{\begin{array}{l}\dfrac{1}{N_\text{A}}\displaystyle\sum_{i=1}^{N_\text{A}}(p_\text{i} - \text{AUC})^2/(N_\text{A}-1) \\ +\dfrac{1}{N_\text{I}}\displaystyle\sum_{i=1}^{N_\text{I}}(q_\text{i} - (1-\text{AUC}))^2/(N_\text{I}-1)\end{array}} \tag{17}$$

The second term is just the variance of the AUC due to the inactives, of number N_I, where we define q_i as the probability inactive i has a higher ranking than a randomly chosen active. If we try to calculate the antiAUC this way we must just get (1 − AUC), since the probability that either an active is more highly ranked than an inactive, or an inactive is more highly ranked than an active has to sum to one. Hence, we can substitute (1 − AUC) for $<q>$. The rest of the form of the standard error of the mean ought to be obvious now, i.e., we are just adding together the squares of the errors due to the finite number of actives, N_A, and inactives, N_I and taking the square root.

Another possibility is simulation. If there are N_A actives and N_I inactives, pick a random set of N_A from the active set, allowing replacement, and similarly choose N_I inactives. Do this many times until you can estimate a 95% confidence range. This avoids the error bars on the AUC crossing zero or one, but does depend on the examples actually being representative. Simulation is not an automatic panacea.

But suppose you don't have the list ranking actives and inactives. Suppose all you have is the AUC and the number of actives and decoys. Can you still say something about the error bars? There are two ways to do this, the first is to assume you know something about the underlying distribution of values of actives and decoys and this approach leads to, for instance, the well-known Hanley formula for AUC error. The insight of Hanley was that if the distribution functions look exponential the mathematics needed to calculate the variance is easy. The resultant formula for the standard error of an AUC is:

$$\text{SEM (AUC)} = \sqrt{\begin{array}{l}\dfrac{\text{AUC}\,(1-\text{AUC})}{N_\text{A}N_\text{I}} + \dfrac{(N_\text{I}-1)\,\text{AUC}^2\,(1-\text{AUC})}{N_\text{A}N_\text{I}\,(1+\text{AUC})} \\ +\dfrac{(N_\text{A}-1)\,\text{AUC}\,(1-\text{AUC})^2}{N_\text{A}N_\text{I}\,(2-\text{AUC})}\end{array}} \tag{18}$$

The first term is always small and represents the "standard" variance for a probability (*see* later in the text) calculated over $N_A \times N_I$ pairs of actives and inactives. The second term captures the contribution to the variance of the actives and the third term, that of the inactives. Although the Hanley result is commonly used I have never seen it derived, even in the original paper! As such, I work it out it in Appendix 3 and show that the Hanley assumption also leads to a nice form for the ROC curve for any AUC value.

Given reasonably large N_i and N_a we get:

$$\text{SEM (AUC)} \approx \sqrt{\frac{\text{AUC}^2\ (1-\text{AUC})}{N_A\ (1+\text{AUC})} + \frac{\text{AUC}\ (1-\text{AUC})^2}{N_I\ (2-\text{AUC})}} \quad (19)$$

The second method is to assume we know nothing about the distributions so as not to bias the end result, we just use the basic definition of the AUC as a probability and average over all combinatorial possibilities that have the right AUC [7]. While this unbiased approach is appealing, it does not seem to give substantially better results than the Hanley formula and is *considerably* more complicated. As such, for this review we shall stick with Hanley's formula.

The algebraic form of the variances for actives and inactives for the Hanley look quite different but are actually trivially related. If we swap $(1 - \text{AUC})$ with AUC then the two terms interchange. This is as it should be because if we redefined an active as an inactive and vice versa we would naturally get an AUC of $(1 - \text{AUC})$, as described previously. As such there is really only one term and the SEM could be rewritten more elegantly as:

$$\begin{aligned}\text{SEM (AUC)} &\approx \sqrt{\frac{\text{AUC}_{\text{active}}^2\ (1-\text{AUC}_{\text{active}})}{N_A\ (1+\text{AUC}_{\text{active}})} + \frac{\text{AUC}_{\text{inactive}}^2\ (1-\text{AUC}_{\text{inactive}})}{N_I\ (1+\text{AUC}_{\text{inactive}})}} \\ &\approx \sqrt{\frac{V\ (\text{AUC}_{\text{active}})}{N_A} + \frac{V\ (\text{AUC}_{\text{inactive}})}{N_I}} \\ V\ (x) &= x^2\ (1-x)/(1+x) \end{aligned} \quad (20)$$

From this elegant formula we can consider the relative contribution to the error from the number of actives and the number of inactives. In retrospective studies we are usually limited in the number of actives we have but tend to feel that there are an infinite number of inactives to play with, e.g., just choose compounds at random. Of course, this is not a good strategy because the performance of a method should be judged in a real world setting, for instance, what might actually get screened or a chemist might

make. In addition, a method might need to be judged as to what it brings relative to other, perhaps simpler methods. Whatever the rationale for choosing decoys we are still left with the question of "how many." Let's rewrite the Hanley formula a little bit and assume that the actual number of decoys is going to be significantly greater than the number of actives,

$$\begin{aligned}\mathrm{SEM(AUC)} &\approx \sqrt{\frac{\mathrm{V(AUC_{active})}}{N_\mathrm{A}}\left(1+\frac{N_\mathrm{A}}{N_\mathrm{I}}\frac{V(\mathrm{AUC_{inactive}})}{V(\mathrm{AUC_{active}})}\right)} \\ &\approx \sqrt{\frac{V(\mathrm{AUC_{active}})}{N_\mathrm{A}}}\left(1+\frac{1}{2}\frac{N_\mathrm{A}}{N_\mathrm{I}}\frac{V(\mathrm{AUC_{inactive}})}{V(\mathrm{AUC_{active}})}\right. \\ &\quad \left.-\frac{1}{8}\left(\frac{N_\mathrm{A}}{N_\mathrm{I}}\frac{V(\mathrm{AUC_{inactive}})}{V(\mathrm{AUC_{active}})}\right)^2\cdots\right) \\ &\approx \sqrt{\frac{V(\mathrm{AUC_{active}})}{N_\mathrm{A}}}\left(1+\frac{1}{2}\frac{N_\mathrm{A}}{N_\mathrm{I}}\frac{V(\mathrm{AUC_{inactive}})}{V(\mathrm{AUC_{active}})}\right)\end{aligned} \tag{21}$$

As such, we can see that the ratio of inactives to actives controls how much greater the error in the AUC is compared to when there are an infinite number of decoys. Let's assume we have a reasonably good method and the AUC is 0.75. The ratio of variances in this equation is then $(7/15) \approx 0.467$. A tenfold excess in decoys means the error is only about 2.3% higher than the limiting result! Even a threefold excess only adds ~8% to the error due to the finite number of actives. Clearly you do not need a lot more inactives than actives to estimate an AUC. Yet despite this, what ratios do we often find in use? The commonly used DUD decoy set [8] uses a ratio of 1:40 that lead to an error only 0.5% higher than the $1{:}\infty$ limit. If you accept the AUC is a good measure of performance (and some do not [9]) and want to calculate it accurately there is really no need to ever have more than a tenfold excess and really threefold is completely adequate in most cases. In fact, if you play with the formula above you find that the ratio of variances goes to zero the better a method becomes, i.e., the higher the AUC, the smaller the difference the number of inactives actually make.

2.5. Deconvoluting Variances

Let's look at this further in the context of measuring the expected AUC for a virtual screening method using the DUD set. DUD is made of 40 targets and the AUC for any method is going to vary from target to target. So if we are interested in an accurate average AUC over all 40 we will have:

$$\begin{aligned}\mathrm{Err}(95\%) &= 2.02\sqrt{\mathrm{Var_{method}}/40} \\ \mathrm{Var_{method}} &= <\mathrm{AUC}^2> - <\mathrm{AUC}>^2\end{aligned} \tag{22}$$

Here, "2.02" is the Student's *t*-value for a sample of 40. Now the variance between targets is going to include contributions from the variance for each target, which as we have just seen can be approximated by the Hanley formula. As we know variances add, we can write down the following:

$$\mathrm{Var}_{\mathrm{method}} = \mathrm{Var}_{\mathrm{intrinsic}} + < \mathrm{Var}_{\mathrm{Actives}} > + < \mathrm{Var}_{\mathrm{Inactives}} > \quad (23)$$

The term $\mathrm{Var}_{\mathrm{intrinsic}}$ is the variance of the computational method if DUD had an infinite number of actives and decoys while the other two terms are the average contributions to the variance from the individual systems that make up DUD due to having a finite number of actives and decoys. Now we know (or can estimate) all the terms in this equation except $\mathrm{Var}_{\mathrm{intrinsic}}$ and so we can actually estimate the true intrinsic variability of our method! If we can identify all the sources of error that we know, and know the observed variability, we can subtract off to get an estimate of the "true" variability.

Following this procedure for a sample of methods leads to estimates as in Table 2. A typical computational method has an intrinsic variance at least an order of magnitude larger than that from the finite number of actives and decoys. With such an estimate of $\mathrm{Var}_{\mathrm{intrinsic}}$, we can turn this around and see just how much of a difference a given number of actives and decoys make to the estimate of the error of the average AUC of a method over DUD. The result is that the contribution from the latter to the confidence limits is at most 5%, i.e., if we had an infinite number of actives and decoys our error bounds would be 0.95 of what they currently are. And of this 5%, how much of this is due to the number of decoys, recalling that DUD has a 40-fold excess of decoys to actives? About 1% of that 5%, i.e., 0.05%! Even using

Table 2
The contributions to the variance from either the finite number of actives and decoys or from the intrinsic variance of various methods over DUD. MACCS keys [34] and LINGOS [35] are both "2D" methods, i.e., only use connection table information

Method	Active + inactive variance	Intrinsic method variance
FRED	0.002	0.023
ROCS	0.002	0.041
MACCS	0.002	0.030
Lingos	0.002	0.035

only a threefold excess of inactives would add less than 1% to the total error of the average AUC.

This can get really absurd. At a meeting on computational methods in Japan (Drug Discovery Informatics Forum, October 8th, 2008, Osaka) contributors were required to submit answers on just two systems for which there were 16 actives and 83,339 decoys for the first system and 307 actives and 83,376 decoys for the second. Not only is $N = 2$ for the calculation of error bounds, but the contributors were forced to perform calculations on several orders of magnitude more decoys than was necessary. This is why, I believe, this article makes sense. We, as a community, should be required to understand basic statistics.

2.6. False False Positives

2.6.1. The Effect of False False Positives on the AUC

But does it matter? We tend to believe computation is "cheap" so perhaps having to run calculations for many times longer than is necessary doesn't matter. But consider the case of false false positives (FFP). If you have a large enough set of decoys it is quite possible that some of them would actually be active. This gets more likely if we have a big enough set of actives or if we have chosen them based on physiochemical properties to be similar to active ligands (as is the case with the DUD decoys). So does this matter? Actually for the AUC it doesn't matter much. Remember the AUC is the probability an active will be scored higher than an inactive. If we have a few active compounds in the set of decoys, these should be scored equivalently to the known actives, i.e., treating them as inactives means our AUC will appear smaller than it should be. If the FFP fraction is f,

$$\mathrm{AUC}_{\mathrm{observed}} = (1 - f) \times \mathrm{AUC}_{\mathrm{true}} + f \times 0.5 \qquad (24)$$

i.e., the weighted average of the probabilities. So

$$\mathrm{AUC}_{\mathrm{true}} = (1/1 - f) \times \mathrm{AUC}_{\mathrm{observed}} - 0.5 \times (f/1 - f) \qquad (25)$$

Typical hit rates in HTS screening are about 0.1%. In fragment based screening, they are of the order of a few percent but only for low affinity leads. Let's be cautious and say that because we've carefully selected the decoys to be like the known actives there is a 1% FFP rate. In this case an AUC of 0.75 would in reality be 0.7525, i.e., 0.3% higher. Given the variances noted above, this difference is really unimportant.

2.6.2. The Effect of False False Positives on Enrichment

However, consider that other popular measure of performance known as enrichment. So far we have concentrated on the AUC because it is a well-understood measure of performance. It also correlates quite strongly with enrichment when considered as an *average* property, i.e., single targets may have an ROC curve where enrichment is poor but the AUC is great or, more usually,

vice versa, but this is rarely observed when you average over many systems. However, enrichment resonates with modelers because it corresponds with the typical use case of a virtual screen, i.e., we use a method to "enrich" a subset of compounds with active molecules, reducing the number of compounds to physically screen. As such, it is perhaps the most common performance metric to be found in papers on virtual screening. Enrichment is typically defined in these papers as the ratio of the number of compounds found in a subset to the number expected, e.g., if we look at the top 1% of compounds we would expect to find 1% of the actives. If instead we find 5% of the actives we have enriched by a factor of five. If all the compounds were active we'd have a maximal enrichment of 100.

We are going to consider the statistics of enrichment in some detail, but first we shall make a case for redefining enrichment. We do this because there is little point developing the statistics of a fundamentally unsound metric, which is what the standard definition of enrichment gives us. To see why, consider the following case study. We have 100 compounds consisting of ten actives and 90 inactives. We look at the top ten compounds and are delighted to see that there are five actives in this subset. By the traditional definition of enrichment we would have expected only one active, but we found five so the enrichment is fivefold. Now, consider if we add another 90 inactive compounds to the mix. Suppose in the original experiment the tenth compound had a score of α. We would expect about the same fraction of the new inactives to pass this threshold as before, i.e., another five should have the score above α. Does this mean the enrichment is still five? We now have ten inactives and five actives in the top 15 compounds. That is 15 out of 190 total, i.e., we are no longer looking at the top 10%, but the top 7.9%. At this percentage we would expect to see 10×0.079 actives, therefore the enrichment at 7.9% for this example is $5/(0.079 \times 10) = 6.3$. What is the enrichment at 10%? We don't know unless we have the original list of actives. It will be less than 6.3, because enrichment falls off as we go further down the list, but it is highly unlikely to still be five. This illustrates that enrichment is a function of the ratio of inactives to actives, let's call it R, not just a function of the method.

But there is an even bigger problem with enrichment. Suppose we go the other way and reduce the number of inactives to be the same as the number of actives, i.e., ten. What is the enrichment at 10%? Well, this is now just the top two compounds. Given that there are equal number of actives and decoys the "no enrichment" expectation is there be one of each. But in this example it is likely the top two will both be actives. This means the enrichment would be two. But wait, we get the best possible result, i.e., all actives in the top 10% but only get an enrichment of two, not ten? When R is small you can only get a maximum enrichment equal to the ratio

of the total population to the number of actives, i.e., $1 + R$. In our first example, where we had ten actives and 90 inactives, i.e., $R = 9$. This means the maximum enrichment is ten. We looked at the top 10% of this set so we could still see the expected maximum enrichment. Had we looked at the top 5%, where the maximum possible enrichment should be 20, we could only have seen at most an enrichment of ten. Ajay Jain refers to this as the "saturation" effect [10]. If we take this to the limit of nearly all actives, the enrichment can never make it above 1.0. If the pan you are using to sift for gold contains mostly gold, you can't enrich very much.

The usual response to both these criticisms is to include "an excess of inactives." If R is very high then the enrichment plateaus at a limiting value and the saturation effect disappears. However, it is very typical to see enrichments at 1%, where the maximum enrichment is 100 and R is much less than 99. For instance, in the DUD dataset $R = 40$ so that even at 2% we are working with a reduced dynamic range. Furthermore, even if we have a full dynamic range, the measured enrichment may depend on R in the way discussed above. And what is worse, these effects are more pronounced the better the method. Contrast this with the calculation of the AUC. The AUC is independent of R, by definition, and even the error in the AUC is only weakly dependent on R once R is greater than about three. In addition, this dependency gets weaker the better the method. Can we apply these lessons to enrichment?

The alternative proposed by Jain is called the ROC enrichment. For this we simply replace the meaning of "top x%" to mean "top x%" of inactives. Then the enrichment is the ratio of the fraction of actives to the given fraction of inactives, i.e., (y/x) for the points on the ROC curve. In the above example this would be the fraction of actives with a score above α divided by the fraction of inactives with a score above α. This ratio does not change when the number of inactives is increased. There is no asymptotic approach to a "real" value as R becomes large, it is independent of R. Similarly, there is no saturation effect. No matter what the number of actives we can still see the maximum expected enrichment. If the method actually has no information, i.e., is random, then we get an ideal ROC curve that goes from the point $(0, 0)$ to $(1, 1)$, i.e., the 45° incline of slope one, i.e., the enrichment at all points is one. Of course, the error estimates of this quantity will definitely depend on R and that is what we shall consider shortly.

But first, let's look at the stability of ROC enrichment with respect to a nonzero FFP rate. Suppose we have 5% of the actives higher than 1% of the inactives in our list, i.e., we have a ROC enrichment of five. If we have our FFP rate of 1% this means that actually we were looking at an inactive % of roughly 0.95% not 1% because about 5% of those inactives were actually FFPs.

Essentially, we have to be further to the right on the ROC curve than we expected to find the percentage of actives higher than 1% of the inactives, i.e., we look at about 1/0.95%, i.e., about 1.05%. How much greater is the percentage of actives at 1.05%? Well, that depends on the slope of the ROC curve, as illustrated in Fig. 4. This means we can't know for sure unless we have the curve in front of us. However, we can put a lower bound on things by assuming the slope is greater than one. This is not true past a certain point for the ROC curve as illustrated in Fig. 5, but is usually true for reasonably good methods in the area of concern. Actually, the slope of the ROC curve is related to the information content at that point but that is a digression, if an interesting one. We can also guess an upper bound to the slope because it is likely to be less than the enrichment, i.e., the slope of a ROC curve usually is a monotonically decreasing function, i.e., it gets smaller as the fraction of inactives increases. Therefore, the line from (0, 0) to the ROC enrichment point, here (0.01, 0.05), will have a greater slope than the ROC curve, i.e., be an upper bound. The larger the enrichment the better the approximation becomes, as illustrated in Fig. 5. Putting these together, a *lower bound* on the enrichment error is 1%. Compare this to the 0.3% error in the AUC with the same 1% FFP rate. An upper bound on the slope for our example is the ROC enrichment, leading to an upper, and more realistic, bound of a 5% error. This means that the actual error in the enrichment at 1% due to a 1% FFP is somewhere

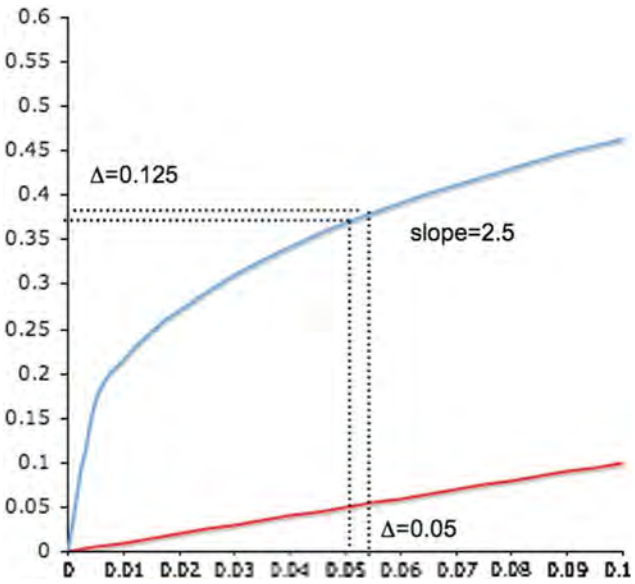


Fig. 4. Illustration of the slope amplification of an error in the *x*-coordinate due to a small fraction of FFPs. (The slope really is 2.5, it is the different scales on the *x* and *y* axis that makes it look less!).

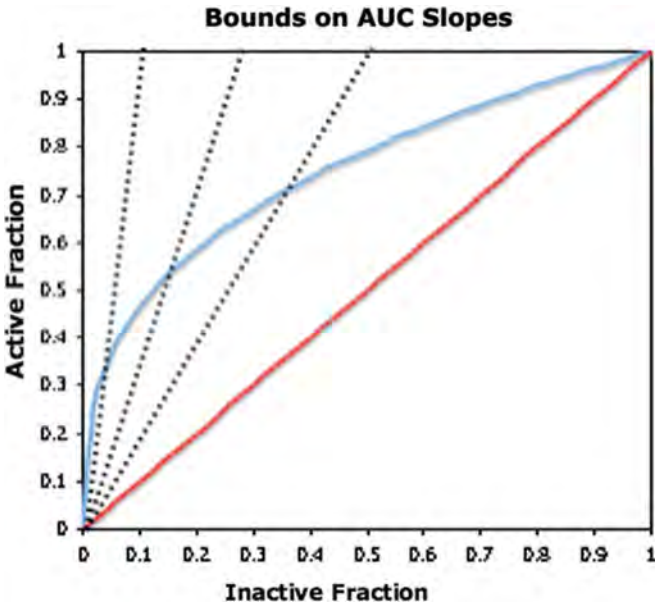


Fig. 5. In most areas of interest ($x = 0$ to 0.2) the slope of the *Upper Curve* (AUC $= 0.75$), lies between that of the *Lower Curve* and the slope of the line from the *Origin*.

between 1 and 5%. The general formula for enrichment error is as follows:

$$
\begin{aligned}
\mathrm{Err} &= (E_{\mathrm{actual}} - E_{\mathrm{apparent}})/E_{\mathrm{apparent}} \\
E_{\mathrm{apparent}} &= A_{\mathrm{apparent}}/I_{\mathrm{apparent}} \\
E_{\mathrm{actual}} &= A_{\mathrm{actual}}/I_{\mathrm{apparent}} \\
I_{\mathrm{actual}} - I_{\mathrm{actual}} E_{\mathrm{apparent}} F &= I_{\mathrm{apparant}} \\
I_{\mathrm{actual}} &= I_{\mathrm{apparant}}/(1 - E_{\mathrm{apparent}} F) \\
A_{\mathrm{actual}} &= A_{\mathrm{apparent}} + \left(\frac{\partial A}{\partial I}\right)_{A=A_{\mathrm{actual}}} (I_{\mathrm{actual}} - I_{\mathrm{apparant}}) \\
&= A_{\mathrm{apparent}} + \left(\frac{\partial A}{\partial I}\right)_{A=A_{\mathrm{actual}}} I_{\mathrm{apparant}} E_{\mathrm{apparent}} F/(1 - E_{\mathrm{apparent}} F) \\
E_{\mathrm{actual}} &= E_{\mathrm{apparent}} + \left(\frac{\partial A}{\partial I}\right)_{A=A_{\mathrm{actual}}} E_{\mathrm{apparent}} F/(1 - E_{\mathrm{apparent}} F) \\
\mathrm{Err} &= \left(\frac{\partial A}{\partial I}\right)_{A=A_{\mathrm{actual}}} F/(1 - E_{\mathrm{apparent}} F)
\end{aligned}
\tag{26}
$$

where F is the FFP percentage, I is the fraction of inactives, A the fraction of actives, and E is the enrichment. The slope term set to the lower bound of one gives a error equal to the FFP rate; set to

the upper bound of the enrichment it leads to an error bound slightly larger than the FFP rate multiplied by the enrichment. As such, the error percentage of the enrichment will depend on performance. Not only does the maximum error grow with the enrichment (and the FFP rate) but the slope of the ROC curve will approach its upper bound, i.e., an enrichment of ten at 1% is much more likely to have a FFP derived error of 10% than 1%. The error in the AUC also increases with performance but only slightly; an AUC of 0.95 has an error due to a 1% FFP of only 0.5%.

What we have just illustrated is that enrichment is quite sensitive to the FFP rate, certainly compared to the AUC. This suggests that it would be well worth making sure that the decoys really are decoys. This, then, is the real value of knowing you do not have to have a large number of decoys – better to choose a few good ones rather than a lot of potentially lousy ones. Of course, there is one element left here to be considered – what are the error bounds on enrichment? The astute reader will see the potential flaw in assuming a reduced R will be as innocuous to the enrichment as it is to the AUC because of the magnifying effect of the slope of the AUC curve, and indeed I will shortly show this to be the case.

2.7. The Variance of Fractions and Probabilities

The first thing we need is an estimate the potential error in either the fraction of the inactives or the actives, i.e., the x or y values of a point on the ROC curve. To get to this we can introduce one of the most useful tricks in statistics – the variance of a fraction. Suppose we want to know what fraction of a population has a particular property. We could do this by inventing a characteristic, say χ that is one if a member has this property and zero if not. We call χ a "characteristic" function. Then the fraction f is just the average of χ over the population, i.e.,

$$f = \frac{1}{N}\sum_{i=1}^{N}\chi_i \tag{27}$$

Now we know a variance can be written as:

$$\mathrm{Var}(f) = <\chi^2> - <\chi>^2 \tag{28}$$

However, we have just said that any χ is only 1 or 0; therefore, any χ^2 can only be 1 or 0, i.e., $\chi^2 = \chi$. Therefore, the variance of a fraction becomes:

$$\begin{aligned}\mathrm{Var}(f) &= <\chi> - <\chi>^2 \\ &= <\chi>(1 - <\chi>)\end{aligned} \tag{29}$$

Since a probability is really just a fraction of events, this lovely result can be used whenever the quantity of interest is probabilistic. As an example, suppose we want to know the fraction

of molecules that have a chiral center. We look at a file of a 1,000 molecules and 600 have such. The best estimate of the probability is 0.6, but with what 95% error bars? The variance is $0.6 - 0.6 \times 0.6 = 0.24$, the sample number is 1,000; therefore, the 95% confidence limits are $\pm 1.96 \times \sqrt{(0.24/1{,}000)} \approx 0.03$. Simple! We can turn this around and say that, if the probability of having a chiral center is 0.6 then the expected range of observed examples is $600 \pm 1{,}000 \times 1.96 \times \sqrt{(0.24/1{,}000)}$, i.e., [570, 630].

Often a probability, p, is small, for instance, the probability of a boron atom in a drug molecule. In this case the variance is approximately just p, as p^2 is so small. In such a case, if the sample number is N so that the number of observations is $M = p \times N$, a rough rule of thumb for the range of likely observed values is:

$$\begin{aligned} \text{range} &= M \pm 2N\sqrt{p/N} \\ &= M \pm 2\sqrt{N \times p} \\ &= M \pm 2\sqrt{M} \end{aligned} \tag{30}$$

Note that p has to be small for this to hold. If applied to our previous example of chiral centers this would have over-estimated the error bounds considerably (± 50 rather than ± 30). We should also remember the warning note sounded previously as to the problems of applying the assumptions of Gaussian *pdf*s away from the predicted average. In this formula $2\sqrt{M}$ is bigger than M whenever $M \leq 4$. Therefore, we have to have at least five observations to apply these bounds (and even then they will be shaky!). Better to use the *logit* transform method as described above and in Appendix 1.

2.7.1. Applied to Enrichment

We now apply this nice result to enrichment. The two quantities that are needed are the fraction of actives when we have seen a fraction of inactives. If all we had to worry about was the fraction of actives this would now be easy. If the fraction of actives if f_a then the error bounds would be:

$$f_a \pm 1.96\sqrt{f_a(1 - f_a)/N} \tag{31}$$

This would translate into a variation in ROC enrichment, E, of:

$$E(f_i) = \frac{f_a \pm 1.96\sqrt{(f_a(1 - f_a)/N}}{f_i} \tag{32}$$

where f_i, the fraction of inactives, is the enrichment identifier, e.g., 1%, 2%, etc. However, as we have seen, an error in the fraction of inactives can also change the value of the enrichment. With our FFP example this was a unidirectional change. Here, we have to consider what the effect of a natural variation in each direction of the fraction of inactives. We follow the normal prescription for

adding variances, except we have to make allowances for the fact the slope of the ROC curve amplifies the error from the inactives. The resultant formula is:

$$\mathrm{Var}(E(f_\mathrm{i})) \cong \frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}} + s(f_\mathrm{i})^2 \frac{f_\mathrm{i}(1-f_\mathrm{i})}{N_\mathrm{i}} \tag{33}$$

Here, $s(f_\mathrm{i})$ is the slope of the ROC curve at inactive fraction f_i. The slope comes in squared precisely because we have to consider f_a as a function of f_i and hence we have to use the rule for transforming variances by the square of $(\mathrm{d}f_\mathrm{a}/\mathrm{d}f_\mathrm{i})$, i.e., the slope of the ROC curve at f_i (Equation 3).

This formula is not so different from that for the variance of the AUC, i.e., we can rewrite so:

$$\begin{aligned}
\mathrm{SEM}(f_\mathrm{a}) &= \sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}} + s(f_\mathrm{i})^2 \frac{f_\mathrm{i}(1-f_\mathrm{i})}{N_\mathrm{i}}} \\
&= \sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}}} \sqrt{1 + s(f_\mathrm{i})^2 \frac{N_\mathrm{a} f_\mathrm{i}(1-f_\mathrm{i})}{N_\mathrm{i} f_\mathrm{a}(1-f_\mathrm{a})}} \\
&\approx \sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}}} \left(1 + \frac{1}{2} s(f_\mathrm{i})^2 \frac{N_\mathrm{a} f_\mathrm{i}(1-f_\mathrm{i})}{N_\mathrm{i} f_\mathrm{a}(1-f_\mathrm{a})}\right)
\end{aligned} \tag{34}$$

If we note that $N_\mathrm{i}/N_\mathrm{a} = R$ and $f_\mathrm{a}/f_\mathrm{i}$ is actually the enrichment, E, then we have:

$$\begin{aligned}
\mathrm{SEM}(f_\mathrm{a}) &\approx \sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}}} \left(1 + \frac{1}{2} s(f_\mathrm{i})^2 \frac{1}{RE} \frac{(1-f_\mathrm{i})}{(1-f_\mathrm{a})}\right) \\
&\approx \sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}}} \left(1 + \frac{1}{2} s(f_\mathrm{i})^2 \frac{1}{RE}\right)
\end{aligned} \tag{35}$$

The last approximation only holds if f_a and f_i are both small. We can be sure this is true for f_i because we typically chose it to be so. The case for f_a small is not so clear-cut because if the enrichment is really good it has to become large. However, if we do use this approximation we get the following nice result. Recall that a reasonable upper bound on the slope is simply the enrichment, E. If we put this into the last equation we get:

$$\mathrm{SEM}(f_\mathrm{a}) \approx \mathrm{SEM}(f_\mathrm{a} | R = \infty) \left(1 + \frac{1}{2} \frac{E}{R}\right) \tag{36}$$

i.e., the increase in the error bounds on the enrichment because R is not infinite, is just $E/2R$. The similar equation for the error bounds behavior of the AUC with respect to R, with AUC $= 0.75$, was about $0.2/R$, i.e., the sensitivity to R is about $2.5E$ higher for enrichment. If we are lucky enough to have an enrichment of ten, often reported at the 1% level, this translates to a 25-fold greater

sensitivity, with respect to R. If $R = 10$, this translates to about a 50% increase in error bounds for the enrichment, compared to about 2% for the AUC. If $R = 40$, i.e., at the DUD level, the percent added to the error is only about 10%, illustrating that the DUD curators clearly knew just what they were doing!

I hope the above illustrates that enrichment is a significantly more difficult quantity to pin down. The beauty of the AUC is that it is much more robust, to both the ratio of actives to inactives and also to the appearance of FFPs as well as being a good estimator of likely performance early in a screen. Why is this? Simply because the AUC derives from information from all actives and decoys, not just a small, and hence statistically suspect, subset.

2.8. Further Applications of Error Propagation

2.8.1. Modelers as a Source of Error

Most methods in complex fields like molecular modeling have multiple sources of error. In a simulation it could be the partial charges, the force-field, incomplete sampling, etc. In docking the list is endless. If we happen to know the effect of one source of error, we can use the above formula to extrapolate to what the error would be if we could resolve completely that component. I ran into this in considering virtual screening results where there is considerable variation in performance from target to target, for instance, over the DUD dataset [8]. The performance of a method on each target is itself uncertain because there are only a finite number of actives and decoys. However, we can estimate this error on a system-by-system basis. As a result we can subtract this system variance from the total variance seen over all the methods and arrive at an estimate of what the variance would be in the limit of an infinite number of actives and decoys. This is a more fundamental characteristic of a method and comes about because we understand how errors add.

Now consider a more controversial application: that of operator error, or variation, in using a program. Suppose we could have a group of people run the same procedure but each does it his or her own way. There would naturally be a variation in outcomes. We have seen this in our SAMPL events [11] where even people using the same version of the same docking program can get quite different results. Suppose our set of modelers tackles the hard problem of evaluating the method over a set of targets. The variation in targets itself is going to provide variability in the average performance, but we can also now use the known user variability to subtract off from the observed variation to get the "natural" variation of the method. Perhaps this is a solution to the well-known problem that plagues methods comparison, i.e., that it depends on the user as to how well a method will work. This is a contentious statement because some will say, with some validity, that the value of a method should be assessed as with the best modeler, i.e., the best-foot-forward argument. However, there are two possible flaws to this argument. The first is that you might not

be lucky enough to have a great modeler (or he or she does not have time for your project), and secondly perhaps there really isn't such a thing as a "good" modeler (this is the contentious bit), i.e., is the variability we see in usage essentially random with a variance we can subtract, or are some modelers really better than others? This sounds like something statistics should be able to tell us, if we had the data.

2.8.2. The "File-Drawer" Effect

There is another area where the concept of modelers as sources of error has application, that of publication bias. As members of a field we are all aware of the problem: only papers with positive results tend to get published and so what we actually read is a biased sample of what is actually done. In the statistics world this goes by the quaint name of the "File-Drawer Effect", i.e., can you account for all the reports of negative results that get shoved into a file cabinet never to see the light of day [12]. The basic approach taken is to look at the size of an effect published and ask the question, "How many negative reports that would not get published would it take to make this positive report look like statistical noise?" If this number is outlandishly large then assume there really is something to the report. If, however, you suspect there really are an outlandishly large number of negative results out there, for instance, you know many people have tried and failed to produce this result, or if the statistics of the positive report were weak, i.e., it would not take many unpublished reports to contradict the finding, then you have cause to suspect the result. *See* Scargle [12] for more details. The approach has developed in the medical literature where the "curse of $p = 0.05$" is prevalent [13], i.e., how many reports might not have been sent to journals because the significance level was slightly less than the generally accepted 95% confidence limit laid down by Fisher and his cabbages? This is an example of what is referred to as "meta-analysis", i.e., the statistical analysis of multiple reports. The classic example is the link between smoking and lung cancer. It was not a single report that convinced the medical community, it was the meta-analysis over an ever-increasing body of evidence. The File-Drawer effect is taken seriously enough that the U.S. Government has attempted to force the publication of all clinical trials, successful or not, to circumvent the problem.

Quantitative application of general meta-analysis to modeling might be hard because, as a field, don't yet have the required statistical sophistication in our reports. However, application of the File-Drawer approach is still possible. Consider whether molecular docking is useful for finding new and novel leads. Sure there are some reports of successes but those are typically by those looking for the effect, e.g., developers of docking programs. Almost any approach that searches for active molecules will sometimes have success, even if by random chance. Any single

report might make claims as to how exceedingly unlikely their finds were by chance, claims usually to be taken with the proverbial grain of salt, but how many nonreports are out there? After all, many, many researchers use docking programs, so where are the overwhelming flood of such papers? There are reports, and also a general consensus, that the scoring functions used by such programs can not rank order molecules by affinity [14, 15], so is it not possible that any success stories come instead from the mass action of many sold on the promise of molecular docking, not its statistically-based reality? I merely conjecture.

2.8.3. Variance-Weighted Averages

Here's a final use of the variance from multiple sources, plus a warning on its application. Suppose we have a series of measurements of a quantity x but we know that some measurements are more accurate than others. It would make sense to trust the accurate measurements more than the inaccurate ones, right? In fact this is exactly right and you can show that a weighed average of the measurements is actually a better average:

$$\mu_{\text{variance weighted}} = \frac{\sum_{i=1}^{N} \mu_i/\sigma_i^2}{\sum_{i=1}^{N} 1/\sigma_i^2} \tag{37}$$

This form arises naturally from considering the product of N different CLT-derived Gaussians. Values with small variances are weighted much more than those with large variances.

As an example of how we might use this, suppose we looked at DUD again and observed that not all the targets have equal numbers of actives. As such, we know the variance of the calculated AUC for each target will likely be different. So perhaps a better way to get the average AUC of a method across DUD is to use the above formula, i.e., give more credence to those targets with less error. *This is incorrect*! To see why, consider a limiting case. Suppose one of those targets has a billion actives (and 40 billion inactives!) The variance of this one target's AUC would now be essentially zero. Therefore, that AUC value would dominate the variance-weighted mean. Should it? Clearly not, because it is just one, admittedly very precisely characterized, system. The AUC of this system does not determine what the mean AUC is over all possible systems, it just tells us about that one system. The fallacy here is that the means, i.e., the AUCs going into the sum are *not* the same properties, i.e., they would not be expected to converge to some mean value just because we can measure them more precisely. Now, suppose we had the same set of DUD targets and a completely different set of actives and decoys for each target and ten times as many of such. When we measure the average AUC for a method on this second set we

would trust it would be more accurate. To combine the average AUC from the first set of calculations with the second we would be perfectly correct to use a variance-weighted scheme because the underlying quantity, the average AUC over DUD, is the same quantity in both cases.

The reason this seems confusing is that we could imagine another limiting case, i.e., where one measurement is very noisy – don't we want to weight this value differently? Consider including a new target into DUD for which we have very few actives. The error in measuring the AUC of this target may be large. By the propagation of errors, we know this system may significantly increase the sum of the errors from each system. On the other hand, we have increased the number of systems, reducing the SEM for this expanded set. We have to weigh the advantage of adding a noisy system against the reduction in the variance over the entire set because there are more systems. This is the consequence of variance for a noisy system – we have to recalculate the average error. Variance weighting only applies when we are averaging the same underlying quantity. When we are averaging quantities we know to be different, we do not variance-weight, we just accept the error may or may not increase, something the basic rules of error propagation allows us to estimate.

2.9. Differences in Results

2.9.1. Independent Samples

If we are presented with two results and two sets of error bars for methods A and B, how do we assess which is better? First of all, if A appears better than B and the tests have been fair, i.e., B was not challenged with a harder test than A (not that this *ever* happens in modeling comparisons!), then the best bet is that A is actually better than B. All statistics can tell you is how good a bet that is. One typical rule that is actually wrong is that if the error bars overlap then the fact A is better than B is not significant, i.e., you would lose the bet that A is better than B 5% of the time. It is wrong both because you don't have to be 95% sure before you might trust something, but also because this is not how errors add. As we have seen, variances add, not errors. As such, a better rule of thumb is that the "joint" error bar i.e., the distance between the estimates of A and B ought to be $\sqrt{2}$ times the individual error bars, assuming they are roughly the same size. This follows because the total variance is the sum of the variances, i.e., twice as big, and the error is derived from the square root of the variance. The proper formula for the joint error bar is almost as simple:

$$\mathrm{SEM}_{\mathrm{joint}} = \sqrt{\frac{\mathrm{Var}_{\mathrm{A}}}{N_{\mathrm{A}}} + \frac{\mathrm{Var}_{\mathrm{B}}}{N_{\mathrm{B}}}} \tag{38}$$

Note that this gives lie to the rule that variances add because we really mean it is the sum of the variances normalized by the sample size; the variances only really add when the sample size is the same

in both cases. Sample weighted variances add. One consequence of this simple rule is that the joint error bar can never be smaller than the larger of the error bars and never bigger than $\sqrt{2}$ times than this same quantity. As such, a more cautious rule is that if one *value* falls within the error bar of the other *value* the result is not significant to the level of the error bars, typically 95%.

Here is an example. There was a recent report of an AIDS vaccine trial that caused a lot of excitement [16]. Of 8,197 recipients of the vaccine, 51 contracted AIDS, compared to 74 of the 8,198 who did not. Should there be excitement? Well, the first reason for hope is that the best guess, given this information and assuming the vaccinated were chosen without any bias, is that 31% fewer people got AIDS who were vaccinated. But how good a bet is this? As we have seen, the variance of a fraction, when that fraction is small, is approximately equal to that fraction. Therefore, the error for the difference in the fractions is:

$$\begin{aligned} \mathrm{Err}(95\%) &= 1.96\sqrt{\frac{51}{8197}\frac{1}{8197}+\frac{74}{8198}\frac{1}{8198}} \\ &= 0.0027 \end{aligned} \tag{39}$$

The difference in fractions, Δf is:

$$\Delta f = \frac{51}{8197} - \frac{74}{8198} = 0.0028 \tag{40}$$

So actually, despite the great fanfare, this result actually falls just outside the normal $p = 0.05$ probability that this is a chance result. However, as we have noted, $p = 0.05$ is a very arbitrary measure and given the complete lack of any statistical signal in previous trials it was easy to see the excitement.

The observant at this point might point out that we have been using the Student's t-value of 1.96 for these combined results. In the AIDS vaccine test there were over 8,000 samples so we can be pretty much assured that is the right value. But what if we are comparing two results with much lower sampling numbers – which number do we use, N_A or N_B or some combination of both? There is a formula for that, of course, derived independently by B. L. Welch [17] and F. E. Satterhwaite [18] in 1946 and 1947, respectively, hence known as the Welch–Satterhwaite equation. It is more general than combining just two variances, allowing an estimation of the "effective" number of observations, n_combined, for any combinations of factors having different variances and degrees of freedom:

$$n_\mathrm{combined} \approx \frac{\left(\sum_{i=1}^{N}\frac{\mathrm{var}_i}{n_i}\right)^2}{\sum_{i=1}^{N}\left(\frac{\mathrm{var}_i}{n_i}\right)^2\frac{1}{n_i-1}} \tag{41}$$

This is essentially just a standard error weighted average of the degrees of freedom. If you stare at it long enough you should be able to convince yourself that it favors the smaller of the degrees of freedom and, in fact, the rule of thumb is to just use this smallest number. Finally, remember it is $(n - 1)$ that goes into the Student's t-test. This is usually a small correction anyway, but it is good to know this has been worked out. That's the thing about classical statistics; it is like the English countryside, made pleasant by the efforts of many over a long period of time.

2.9.2. Correlated Samples

Let's consider another example. Figure 6 shows a hypothetical example of the performance of two methods, A and B across 12 targets. Figure 7 shows the averages of A and B with 95% confidence limits. If we follow the rule of thumb from above clearly there is no statistically significant difference between A and B, i.e., both lie within the error bars of the other. And yet if we look at Fig. 6 again we might notice that for each and every example A is

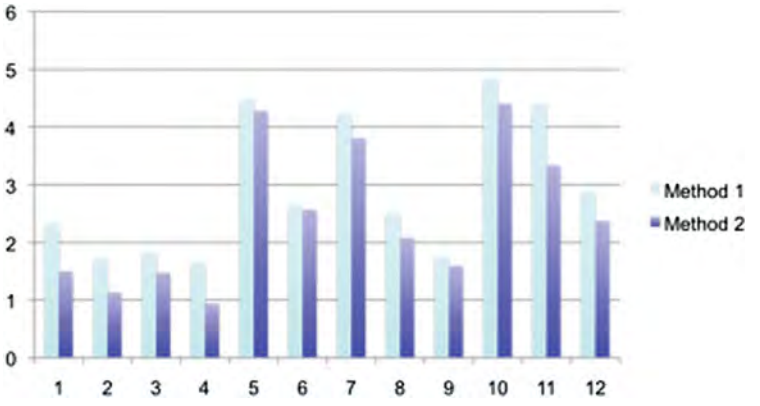


Fig. 6. Methods 1 and 2 for some arbitrary measure. Note that method 1 is always slightly better than method 2.

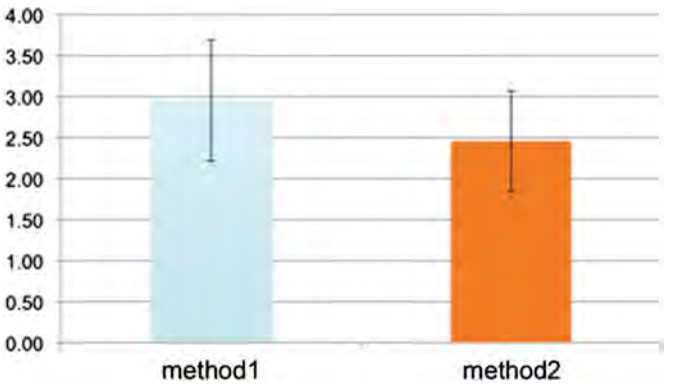


Fig. 7. If we calculate the averages of each method and plot the 95% error bars we see neither method is statistically different from the other.

better than B. If we had to guess what a 13th system would yield we would clearly expect A to be better than B. So what has gone wrong with our analysis?

The key here is to instead look at the difference in performance of A and B, i.e., $C = A - B$. This is shown in Fig. 8. If we calculate the standard deviation of this difference (0.29), and hence the 95% confidence limits on the average difference (0.18) we find that average difference is 0.49 ± 0.18. We can now ask what the probability might be that, despite appearances, B is actually better than A. Note first that this is a one-sided test. We are not asking if in fact A and B are in fact equivalent and the variation is actually due to chance, which would be the two-sided test. We want to know if B is really better than A. For that we just have to evaluate the area under a Gaussian centered at 0.49, with width calculated from a standard deviation of 0.29 and a sample size of 12, that lies below 0.0, i.e., that would correspond to B being better than A. Let's remind ourselves of how we do this. We have the *erf* function describing the area under a Gaussian from zero to x:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{42}$$

This function goes to one when x becomes large, i.e., it measures the area from zero to some positive value. However,

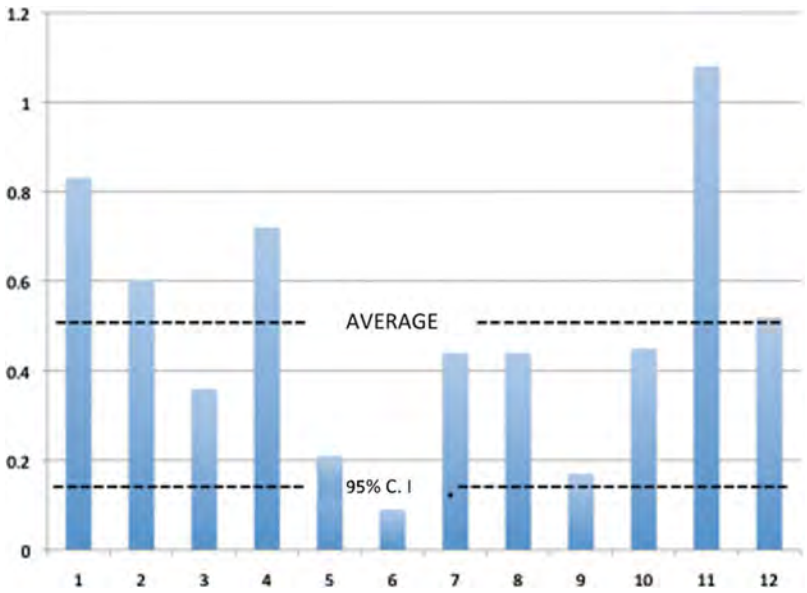


Fig. 8. Differences between methods 1 and 2, plus the average of the differences and the 95% confidence interval on this quantity. Clearly, the average difference is way beyond the 95% confidence interval, indicating the difference is statistically significant.

want the fraction of area beyond x and we also want to account the area from negative infinity to zero, i.e., one-half of the probability density. So:

$$\begin{aligned} p &= 1 - 1/2 - 1/2\,erf(x) \\ &= 1/2(1 - erf(x)) \end{aligned} \tag{43}$$

We have the CLT telling us that the average is distributed as:

$$pdf(x) = \sqrt{\frac{N}{2\pi\sigma^2}} e^{-N(x-\mu)^2/2\sigma^2} \tag{44}$$

We change variables from x to t by adding the mean, μ, and scaling by $\sqrt{2}\sigma/\sqrt{N}$, i.e., we want the *erf* function from $x = 0$ to $x = 0.49/(0.29\sqrt{2}/\sqrt{12})$, i.e., $x = 4.14$. As we have *erf* $(4.14) = 0.99996$, we arrive at $p = 0.00002$. This is a very small number but this should not be surprising, as the average difference between the two methods is almost three times the 95% confidence limits for that difference.

Note, we could also ask a different question: what is the probability of B being better than A on a 13th system. This asks how much probability density of the difference lies below 0.0 for the *pdf* of a single measurement. We would like to say this could be calculated from a Gaussian centered at 0.49, with a standard deviation of 0.29 but where the sample size is one. This would give a result of $p = 0.11$. However, this assumes the *pdf* for a single measurement is Gaussian. There are tests for this, i.e., whether the *pdf* is in fact Gaussian. In fact there are several tests, for example the Kolomogorov–Smirnov test, the Anderson–Darling test, or the Shapiro–Wilk test. These tests are fairly mathematical and I won't say more about them other than that they are referred to as tests for "normality." They are called this because seeing Gaussians is so normal statisticians refer to the Gaussian as the "normal form". What I will say is that assuming the differences of quantities follows a normal form is often a reasonable assumption, even if the *pdf* for each type of measurement is *not* normal. This is a common observation. Some turn to nonparametric statistics precisely because they are concerned as to the nonGaussian nature of their distributions, but if they are actually interested in the differences of quantities classical statistics may still apply.

Now let's think about why we get such a different result when we consider the differences of methods. What has happened is that in our first analysis we forgot the magic incantation "i.i.d." The systems used to test A and B are not independent, they are the same systems. This means that the results between A and B are *correlated*. In all the examples we have considered so far we have assumed that two approaches have been tested on randomly chosen systems and that the means and variances reflect this. For example, in the case of the AIDS vaccine the two sets of volunteers

are different sets of people. If in our above example there were 12 results for A on one set of targets and 12 results from B from a different set of targets that led to Fig. 7, then we would be perfectly correct to say there was not much difference in the methods. It is only because we see the results line up on the *same* targets that we know A must be much better than B.

Let's look at the mathematics of this a little more closely. First, the difference between the mean score of the two methods is the same in either analysis; therefore, it is only how we perceive this difference in relation to the variance. If we were treating these as results from independent trials we know the variance would look like:

$$\sigma^2_{\text{Independent},A-B} = \frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i^A-\mu_A)^2+\sum_{i=1}^{N}(x_i^B-\mu_B)^2\right) \quad (45)$$

And 95% error bounds would be:

$$\text{Err}(\text{independent},95\%)_{A-B} = 2.2\sqrt{\sigma^2_{\text{independent},A-B}/N} = 0.99 \quad (46)$$

Here, the value of 2.2 is the value of the Student's t-distribution for 11 degrees of freedom. As the difference in the averages for A and B is 0.49 it should be clear why the two methods look indistinguishable if the trials are independent. Now consider this as a set of 12 differences. Then, instead, the variance would look like:

$$\sigma^2_{\text{dependent},A-B} = \frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i^A-x_i^B-\mu_A+\mu_B)^2\right) \quad (47)$$

And the 95% error bounds would be:

$$\text{Err}(\text{dependent},95\%)_{A-B} = 2.2\sqrt{\sigma^2_{\text{dependent},A-B}/N} \approx 0.18 \quad (48)$$

i.e., we are taking the average of the squared differences of the differences between the means and the differences from each system. We then compare this error bound to the average difference of 0.49 to arrive at our dependent conclusion. Let's expand each term that makes up this variance:

$$\begin{aligned}\sigma^2_{\text{dependent},A-B} &= \frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i^A-\mu_A)^2+\sum_{i=1}^{N}(x_i^B-\mu_B)^2\right.\\ &\quad\left.-2\sum_{i=1}^{N}(x_i^A-\mu_A)(x_i^B-\mu_B)\right)\end{aligned}$$

$$= \sigma^2_{\text{independent},A-B} - \frac{2}{N-1}\sum_{i=1}^{N}(x_i^A - \mu_A)(x_i^B - \mu_B)$$

$$= \sigma^2_{\text{independent},A-B} - 2\text{Cov}(A,B)$$

$$\text{Cov}(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i^A - \mu_A)(x_i^B - \mu_B) \tag{49}$$

We define the second term of line two as the *covariance* of A and B. The covariance tells us if A and B tend to vary in the same way, e.g., if A gets bigger so does B and vice versa. If the two methods are correlated the covariance is positive and acts to reduce the variance. Methods that are related are much more likely to be correlated than those that rely on different assumptions or preconditions, for instance, docking programs are likely to be correlated because they make similar assumptions and use similar data, i.e., the structure of the protein. They are even more likely to be correlated if methods are mere variants of each other, for instance, different versions of the same program. So if a developer of a method wants to know if a change in approach makes a difference on the test set he or she has in hand then correlation has to be taken into account.

Now, it is possible that method A and B would be anticorrelated. In this case, whenever A was better then expected (relative to its mean) B would be worse and vice versa. It is unusual to see anticorrelation, but it can occur. For instance, a docking method might do well when there are multiple ways a ligand can bind, a case where a ligand-based method might do badly. Similarly, a ligand-based method might do well when there is active site flexibility, a situation a docking program might find more challenging. In this case the dependent variance would actually be higher than expected, making it more difficult to distinguish two methods. Of course, anticorrelation is also an indicator of what methods you might want to combine, a subject with its own statistical fascinations that I will not expand upon here.

Hopefully, this helps to underline the need for standard datasets for the field. To really know if progress has been made we have to be able to tell if a difference is real or merely random. From the CLT we know this requires making the variance small and the number of tests large. Given the intrinsic variability of performance of modeling methods, just about the only way to make the variance small is to have correlation reduce the standard error, which means a common data set. Despite its flaws, such as the chemical similarity of many of its active compounds, this underscores the importance of the DUD dataset. Comparing methods by average performance on different test sets, given the high variance of computational approaches, is essentially meaningless.

Of course, adopting common datasets is but one part of a progressive field. There is also the need to recognize that if changes in a method are made specifically to suit that standard set, then the prospective power, i.e., the reliability on as yet unseen systems, is jeopardized. Addressing the issue of such parameterization is a central part of statistical learning, something I also do not have the room to cover but that in many ways is the driving force behind my own commitment to gain a better understanding of statistics.

2.10. The Student's t-test and ANOVA

As a last instance of the power of simple, Gaussian-based, statistics, I want to consider the Student's *t*-test. We have been using the Student's *t*-distribution throughout this article, i.e., whenever we want to calculate a significance level for a small sample, but Gosset developed this function for a slightly different use. If we have a set of measurements on a set of objects that either have or do not have a specific property, then does that property affect the results? For instance, in Gosset's case it might have been whether the taste of Guinness was as magnificent with or without a given strain of yeast.

Given the formulae we already have seen this might seem straightforward. If we calculate the average of the measurements with the property and the average without we can calculate whether this difference is statistically likely or not. However, this is not quite the question posed. We want to know if two populations are actually drawn from the same population, not the difference of two populations already considered different. Put another way, any two samplings from the same population will inevitably have slightly different means and slightly different variances. We can tell if these differences are significant? If we know what the expected distribution of means and variances of subsets of a single population looked like we could then address whether the means and variances of a couple of given subsets look unusual or not.

In this way we are testing a "NULL" hypothesis. Here, the NULL hypothesis is that the two samples are from the same population, i.e., same mean, same variance. We should first disprove this before we then apply our previous formula for the likely difference between the populations. If this has confuses you, join the club! It always confuses me. It gets worse in that there is something called the Welch test that removes from the NULL hypothesis the assumption the variances need be the same. In this case the Welch test is essentially just the likelihood of seeing differences in averages, just as we have analyzed above. Just remember that the Student's *t*-test is what you do to see if there is any effect at all (i.e., disprove the NULL hypothesis). Then we try and estimate what the magnitude of an effect might be.

So, our two samples each have a mean and each have a variance. If the NULL hypothesis is that these are drawn from the same population then the actual mean will be some sort of

average of the two means and the actual variance would be some sort of average of the two variances. As the likelihood of the two means being different is a function of the actual variance, we are interested in the best estimate of the "true" variance from the two sample variances. This is termed the "pooled" variance and looks like:

$$\mathrm{Var}_{\mathrm{pooled}} = \frac{(n_1 - 1)\mathrm{Var}_1 + (n_2 - 1)\mathrm{Var}_2}{n_1 + n_2 - 2} \tag{50}$$

That is, it is a simple weighted average, just like the best estimate of the mean would be, except that we use $(n_i - 1)$ not "n_i" as the weights. We use $(n_i - 1)$ because this is the denominator in the unbiased estimator of the variance and essentially what we are doing is reverting to the straight sum of the squares of the deviations of each measurement from its mean:

$$\begin{aligned}
&(n_1 - 1)\mathrm{Var}_1 + (n_2 - 1)\mathrm{Var}_2 \\
&\quad = (n_1 - 1)\sum_{i=1}^{n_1} \frac{(x_{1,i} - < x_1 >)^2}{(n_1 - 1)} + (n_2 - 1)\sum_{i=1}^{n_2} \frac{(x_{2,i} - < x_2 >)^2}{(n_2 - 1)} \\
&\quad = \sum_{i=1}^{n_1} (x_{1,i} - < x_1 >)^2 + \sum_{i=1}^{n_2} (x_{2,i} - < x_2 >)^2
\end{aligned} \tag{51}$$

Then we divide by $(n_1 + n_2 - 2)$. Why "2"? Because there are now two means, the mean of sample 1 and the mean of sample 2, i.e., the actual degrees of freedom in the sum of squares is two less than the number of measurements.

With the pooled variance in hand we now proceed exactly as if we were indeed calculating the probability that the means are as far apart as we record, i.e. we use the pooled variance to evaluate the probability a couple of samples of the same variance end up with means $<x_1> - <x_2>$ apart. We know this follows the Student's t-distribution where the standard error of the mean will look like:

$$\mathrm{SEM}(< x_1 > - < x_2 >) = \sqrt{\frac{\mathrm{Var}_{\mathrm{pooled}}}{n_1} + \frac{\mathrm{Var}_{\mathrm{pooled}}}{n_2}} \tag{52}$$

i.e., the probability there are actually two distinct populations is the likelihood that a difference in mean of $<x_1> - <x_2>$ or greater is seen. It is likely there are two distinct populations with 95% confidence if:

$$|< x_1 > - < x_2 >| > 1.96 \times \mathrm{SEM}(< x_1 > - < x_2 >) \tag{53}$$

Readers might ask, "But is 1.96 the correct prefactor?" After all, the whole point of this exercise is to assess finite populations." This is what the Student's t-distribution was actually developed for! We also know how to deal with the situation where $n_1 \neq n_2$,

i.e., via the Welch–Satterhwaite equation, i.e., the degrees of freedom is an interpolation between $(n_1 - 1)$ and $(n_2 - 1)$. Because the pooled variance acts as the true variance of both sets of population, Welch–Satterhwaite simplifies as follows:

$$n_{\text{combined}} = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^2}{\left(\frac{1}{n_1^2}\frac{1}{n_1 - 1} + \frac{1}{n_2^2}\frac{1}{n_2 - 1}\right)} \tag{54}$$

The Student's *t*-test is truly a most useful classical result because it addresses that most fundamental of questions: Did something have an effect? It is so prevalent in the medical literature that there are studies of how often it is misused [19]! As a possible example of this, consider the case of the AIDS vaccine considered above; was I correct in calculating the whether the difference in results from vaccination significant or not? Should I not have actually used the pooled variance and the Student's *t*-test? Let's compare the SEM for both:

$$\begin{aligned} \text{var}_1 &= 51/8197 \\ \text{var}_2 &= 74/8198 \\ \text{var}_{\text{pooled}} &= (8196\text{var}_1 + 8197\text{var}_2)/(8196 + 8197) \\ \text{SEM}_{\text{means}} &= 1.96\sqrt{\frac{\text{var}_1}{8197} + \frac{\text{var}_2}{8198}} = 0.0026731 \\ \text{SEM}_{\text{pooled}} &= 1.96\sqrt{\frac{\text{var}_{\text{pooled}}}{8197} + \frac{\text{var}_{\text{pooled}}}{8198}} = 0.0026734 \end{aligned} \tag{55}$$

That is, the results would be essentially identical. It is not hard to see that is if the sample sizes are same for both sets, the two approaches boil down to the same thing. In the general case, it comes down to what NULL hypothesis you are testing against: that there is no difference in those vaccinated and those not, or whether the difference in infection rates between the two groups significantly greater than would be expected by chance.

Before I give an example of using the Student's *t*-test, I should point out that it is really just the initial step towards what is called ANOVA, perhaps the crowning glory of the work of R. A. Fisher. ANOVA stands for ANalysis Of VAriance and is a way to discern if multiple samples belonged to the same distribution. The way it works is by comparing the variance of the possible subsets to the variance of the means of those subsets. To see why this might be an interesting proposition, consider what happens if we calculate a variance but shift the mean. Remember:

$$\text{Var} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - < x >)^2 \tag{56}$$

So if we shift the mean by δ we get:

$$\begin{aligned}\text{Var} &= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - < x > - \delta)^2 \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - < x >)^2 - \sum_{i=1}^{N}2\delta(x_i - < x >) + \sum_{i=1}^{N}\delta^2\right) \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N}(x_i - < x >)^2 + \sum_{i=1}^{N}\delta^2\right)\end{aligned} \tag{57}$$

i.e., the variance gets shifted by δ^2. Suppose we have a large set of measurements for which we calculate the variance. Now break that set up into groups, and calculate the mean of each subset and the variance of each subset with respect to its own mean. We will have the following:

$$\text{Var}_{\text{total}} = \frac{1}{N-1}\left(\sum_{i=1}^{M}(n_i - 1)\text{Var}_i + \sum_{i=1}^{M}n_i\delta_i^2\right) \tag{58}$$

Here, each δ_{i} is the square of the distance of the mean of each subset μ_{i} to the global mean, μ_{total}. This second term is akin to the variance that would be calculated if we only had the global means and the mean of each of the subsets, so we can rewrite the total variance as a sum of two variances, thus:

$$\begin{aligned}\text{Var}_{\text{total}} = &\{\text{Weighted sum of subpopulation variances}\} \\ &+ \{\text{Weighted sum of variance of subpopulation means to the global mean}\}\end{aligned}$$

Fisher reasoned that these two terms ought to be related if, in fact, they were drawn from the same population. This forms the basis of ANOVA. The ratio of the variances of the subsets (the variance "within"), divided by the variance of each subset means to the global mean (the variance "between") gives a test value or "statistic." This statistic is compared to something called the Fisher *F* function. The *F* function is an expected distribution of the ratio of two sums of squares of numbers drawn from normal distributions. It is related to that other function I also don't have the space to describe properly, the χ^2 function, it of the χ^2 test. The χ^2_{N} function is the expected sum of *N* squared values drawn from a normal distribution. The χ^2 test is a wonderful way to look at whether a set of frequencies falls into the expected pattern, assuming a Gaussian *pdf*, and the Fisher function is a ratio of χ^2 functions. I recommend Snedecor and Cochran [20] from my reading list for a gentle introduction to both.

To see the relevance of this to the Student's *t*-test, note that if there are just two subsets, then the first term in the last equation is

simply the expression for the pooled variance, in Fisher's language the *within* variance. The second term is the sum of the difference of each subset mean from the combined mean:

$$\begin{aligned}
\text{2nd term} &= n_1(<x_1> - <x>)^2 + n_2(<x_2> - <x>)^2 \\
&= n_1(<x_1> - 0.5<x_1> - 0.5<x_2>)^2 \\
&\quad + n_2(<x_2> - 0.5<x_1> - 0.5<x_2>)^2 \\
&= n_1(0.5<x_1> - 0.5<x_2>)^2 + n_2(0.5<x_2> - 0.5<x_1>)^2 \\
&= 0.5 \times N(<x_1> - <x_2>)^2
\end{aligned} \tag{59}$$

That is, the *between* variance is just the half the *squared* difference of the means of the two populations, times by the number of samples. In the Student's *t*-test, we take the difference in means and divided it by the standard error, i.e., a square root of a variance, to get our test "statistic." In ANOVA, the Fisher test statistic is the ratio of the variance between the subset means and the sum of variances within the subsets, i.e., just uses the square of the same quantities that go into the Student's *t*-test. Put another way, we can show that, when there are only two subsets, $t^2 = F$.

Here's an example of how I have used the Student's *t*-test in the past. Suppose we want to know when a particular method ought to be applied and when it ought not. Are there criteria that we can use to define "domains" of application? A good case in point is when to use docking. It is going to be harder for docking to work if there are large-scale motions of the protein, but is might also be harder for docking to work if the protein structures we are using are of poor quality. Can we detect this, i.e., can we divine two populations within a large sample of docking studies based on the accuracy of the docking structures. Again, the DUD dataset comes in useful here. Figure 9 shows the distribution of AUC

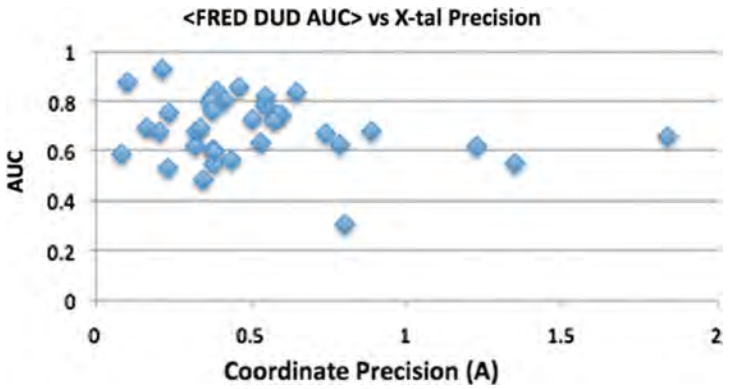


Fig. 9. Distribution of AUC values for the FRED docking program against DUD examples with crystallographic structures against the calculated coordinate precision.

scores for the FRED docking program [21] against the structures in DUD for which coordinate precision can be calculated. Coordinate precision is a global measure of the accuracy of the positions of atoms and is significantly more reliable as a measure of structure quality than the more often quoted "*R*" value. As can be seen, there seems to be a clustering of AUC values on the left (better structures) towards higher values than on the right (poorer structures) in line with our intuition that a docking program ought to care about the quality of the structure it is docking to. If we divide the data into two sets, those structures with coordinate precision greater than 0.7 Å and those less than this value, we find the mean value for the good structures is 0.71 and the mean value for the poor structures is 0.58. Is this difference significant? If we calculate the pooled variance from both sets we find that yes, this is significant at just over the 95% confidence limit.

3. Problems with Statistics

3.1. Overzealousness

The problem with the above analysis is that this is a classic example of using statistics to reinforce a prejudice. If instead of 0.7 Å we had chosen 0.5 Å as our criteria for classification the two subsets are barely different in mean AUC, 0.67 and 0.70, which has no statistical significance at all. In addition, structure quality was not the only criteria I examined for the work. I also looked at the size of the active site pocket, its polarity, the known affinity of the DUD molecules, the size of known ligands, the flexibility of the ligands and their charge. I also looked at the ligand-based program ROCS [22] on the same data set. This means that even if there were a criterion that could be drawn up without first looking at the results, there were in the end about 18 different experiments being tested (Table 3) and so the chances of one of these having a significant result at greater than 95% confidence is actually pretty high. This is, then, an example of the misuse of statistics, or at least the Student's *t*-test, although in my defense I would add that at least six of the characteristics tested appeared significant!

There are standard ways to address this issue. The first, and easiest, is to simply adjust what we view as significant. If we are performing a lot of tests such that the chance of a random result looking significant is high then make the significance threshold harder to pass. The simplest such adjustment is called the "Bonferroni *t*-test." All we do here is to reduce Fishers' $p = 0.05$ value by the number of tests, i.e., in my case the new significance level would become $0.05/18 = 0.0028$, i.e., a result would have to pass the 99.7% significance level. Researchers typically do not use the full Bonferroni test, essentially because it is easy to get to the point where nothing seems significant! Rather, if it is

Table 3
Factors influencing the virtual screening performance over DUD when using either the molecular docking program FRED or the ligand-based method ROCS

Characteristic	Method	
	FRED	ROCS
Size of active site	YES	NO
Polarity of active site	NO	NO
Structure resolution	YES	NO
Ligand size	NO	NO
Ligand complexity	NO	YES
Ligand flexibility	NO	NO
Ligand structure	YES	NO
Ligand charge	NO	NO
Ligand affinity	YES	YES

expected that there are multiple effects of some significance they adopt what are called "step-down" procedures. Here, a *p* value is calculated for each effect and the effects are then ordered by significance. The most significant result has to pass the full Bonferroni correction but the next most significant is only required to pass the Bonferroni correction calculated as if there were one less trial. Each successive test is reduced in severity. When a trial fails its significance test, it and all the less significant trials are assumed not significant, all the more significant ones pass. This is known as the Holm test. There are several variants on it, such as the Student–Newman–Keuls (SNK) test and the Tukey test (a variant of the SNK test). There are also variants on the Holm test that work in reverse order, i.e., start at the least significant result. The advantage of these variants is that they will miss less real effects (i.e., have a lower rate of false negatives) at the cost of occasionally including an effect that was just statistical (i.e., a higher false positive rate). Such methods are often grouped under the name "False Discover Rate" (FDR) control. FDR is becoming increasingly important in technologies such as microarray analysis where many questions are being asked of a single sample of DNA.

3.2. The Nature of Science

This is also an appropriate point to comment on what the role of statistics really should be in science and in molecular modeling in particular. It really should not be used to "prove" things. Ultimately, statistics ought to guide further research. If we see that, statistically speaking, structural precision is important for

docking we should seek out either better or poorer structures and confirm our findings. Furthermore, we ought to work out why this is true, what is the root of the effect we are seeing. This is how we make progress, not simply by observing an effect, declaring the "domain of applicability" and then imagining our work is done. Finally, the goal of science is to measure the size of effects, not simply that there is an effect. Too often a statistical result is "significant/not significant". What science requires is "how much of an effect" because this is really how we develop quantitative science [13]. Rather than merely rate some compounds more worthy of study because they are at the top of some list, we need to be predicting the probability they have the property of interest. Rather than suggesting how something might bind, we need to be able to say with what confidence. Rather than say whether something binds, we need to know how well it binds. Statistics will help, but ultimately we still have to do the science.

4. Further Reading

I have found that perhaps the best source for learning statistics is Wikipedia. Some may be appalled at this but I have repeatedly found it accurate, useful and with important links to other material. And, of course, Google. Where would we be without it?

General-purpose books that I found useful include "Statistical Methods" by George W. Snedecor and William G. Cochran [20]. It is a big book but that is because they go slowly through things in great thoroughness. Stanton Glantz's small book, "Primer of Biostatistics" is a gem [23]. It does not cover as much as Snedecor and Cochran and not in as much depth, but it covers useful things. As does the invaluable, "100 Statistical Tests", by Gopal K. Kanji [24]. I especially like the presentation of nonparametric tests I did not have room to cover. Another good basic book is M. G. Bulmer's "Principles of Statistics" [25]. Its section on the CLT is to be recommended. A bit more advanced is E. S. Keeping's "Introduction to Statistical Inference" [26] and another reliable book is "Statistical Rules of Thumb", by Gerald van Belle [27]. If you really want to know more about ROC curves than 99.9% of scientists, read "The Statistical Evaluation of Medical Tests for Classification and Prediction", by Margaret Sullivan Pepe [28]. It is a bit mathematical in places but hugely comprehensive. An excellent, alternative text on statistics is Phillip I. Good and James W. Hardin's "Common Errors in Statistics (and How to Avoid Them)" [29]. It is mixed bag of observations and recommendations but quite invaluable. There is a nice and simple book on *p*-values from Lemuel A. Moye that I have found useful: "Statistical Reasoning in Medicine" [30]. And for those wishing

to learn more about the limitations of p-values and why testing against the NULL hypothesis is far from true science, look no further than the excellent polemic, "The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives", by Stephen T. Ziliak and Deirdre N McCloskey [13]. For those intrigued by my description of Bayesian statistics, in addition to the Loredo article there is the classic, "Data Analysis: A Bayesian Tutorial" by D. S. Sivia [31]. Many get their start in the Bayesian way of thinking from Sivia's little book. More advanced but more comprehensive is Jean-Michel Marin and Christian P. Robert's "Bayesian Core: A Practical Approach to Computational Bayesian Statistics" [32], and even more so, "Bayes and Empirical Bayes Methods for Data Analysis" by Bradley P. Carlin and Thomas A. Louis [33].

My recommendation is to look before you leap. The most important thing in learning statistics, I believe, is to find a good teacher. As such, find a few books at a level and style with which you are comfortable and can enjoy. I only hope the small selection of methods and ideas I have had the opportunity to present here is half as useful as these books and monographs have proven for me and my work.

Appendix 1: Using *logit* to Get Error Bounds

In the section on enrichment as a metric for virtual screening we arrived at a formula for the variance of the enrichment.

$$\mathrm{SEM}(E(f_\mathrm{i})) \approx \frac{1}{f_\mathrm{i}}\sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}}}\left(1+\frac{1}{2}s(f_\mathrm{i})^2\frac{1}{RE}\frac{(1-f_\mathrm{i})}{(1-f_\mathrm{a})}\right) \quad (60)$$

Let's assume that R, the ratio of inactives to actives, is very large so we just have:

$$\mathrm{SEM}(E(f_\mathrm{i})) = \frac{1}{f_\mathrm{i}}\sqrt{\frac{f_\mathrm{a}(1-f_\mathrm{a})}{N_\mathrm{a}}} \quad (61)$$

In our example, the ROC enrichment was fivefold at 1% inactives because the fraction of actives was 0.05. This would mean the 95% error would be:

$$\begin{aligned}\mathrm{Err}(95\%|f_\mathrm{i}=0.01) &= \pm 1.96\times 100\sqrt{\frac{0.05\times 0.95}{N_\mathrm{a}}} \\ \mathrm{Err}(95\%|f_\mathrm{i}=0.01) &= \pm 40.5/\sqrt{N_\mathrm{a}}\end{aligned} \quad (62)$$

Now the enrichment is 5.0. If $N_a < (40.5/5.0)^2$, i.e., $N_a < 65$, then the lower error bound becomes negative, i.e., a nonsense value. The problem is, as mentioned in the text, that the quantity of interest, the fraction f_a of actives, is bounded between 0 and 1. However, if we transform with the *logit* function it becomes unbounded, just like the Gaussian.

$$y = l(x) = \log\left(\frac{x}{1-x}\right) \tag{63}$$

If we make this transformation then the fraction $f_a = 0.05$ becomes:

$$l(0.05) = \log\left(\frac{0.05}{1-0.05}\right) = -2.944 \tag{64}$$

This is our new mean. Now, we have to recalculate the variance in *logit* space, i.e.,

$$\begin{aligned} \mathrm{var}_l &\approx \left(\frac{1}{< x >(1- < x >)}\right)^2 \mathrm{var}_x \\ &= \left(\frac{1}{0.05(1-0.05)}\right)^2 0.05 \times (1-0.05) \\ &= 21.05 \end{aligned} \tag{65}$$

This means the error bounds on the *logit* version of f_a become:

$$\begin{aligned} l(f_i = 0.01) &= -2.944 \pm 1.96\sqrt{\frac{21.05}{N_a}} \\ &= -2.944 \pm 8.99/\sqrt{N_a} \end{aligned} \tag{66}$$

Suppose we set N_a to a value much less than the "silly" threshold of 65. Let's make it 25. In non-*logit* space this means the 95% range of enrichments is:

$$\begin{aligned} \mathrm{E}(f_i = 0.01) &= [5.0 - 40.5/5.0, 5.0 + 40.5/5.0] \\ &= [-3.1, 13.1] \end{aligned} \tag{67}$$

Clearly, the lower range is nonsense. Now consider the range of f_a in *logit* space:

$$\begin{aligned} f_{a,\mathrm{logit}}(f_i = 0.01) &= [-2.944 - 8.99/5, -2.944 + 8.99/5] \\ &= [-4.8, -1.144] \end{aligned} \tag{68}$$

Now it is perfectly ok that the lower range is negative because *logit* functions go from negative to positive infinity. The final step, then, is to transform these values back to a fraction, using the inverse *logit* function, i.e.,

$$l^{-1}(y) = \frac{1}{1+e^{-y}} \tag{69}$$

And then divide by f_i to get the enrichment. If we do this we get:

$$f_a(f_i = 0.01) = [0.008, 0.247] \quad (70)$$
$$E(0.01) = [0.8, 24.7]$$

Clearly, these are large error bounds, error bounds that actually include an enrichment of less than random! However, they are not negative and they are a reflection of the difficulty of pinning the enrichment down with so few actives. Even if we repeat the analysis with four times as many actives, i.e., $N_a = 100$, the 95% range is still [2.1, 11.5]. The untransformed range for $N_a = 100$ is ~[1.0, 9.0].

Appendix 2: Why Variances Add

Suppose we have two sources of error that can move the measured value away from its true mean, and let's suppose that mean value is zero for simplicity. The CLT tells us that each source alone will produce a distribution of values according to the number of observations and the intrinsic variance of each source:

$$pdf_\alpha(x) = \sqrt{\frac{N}{2\pi\sigma_\alpha^2}} e^{-x^2 N/2\sigma_\alpha^2}; \qquad pdf_\beta(y) = \sqrt{\frac{N}{2\pi\sigma_\beta^2}} e^{-y^2 N/2\sigma_\beta^2} \quad (71)$$

Now x and y are independent variations from the mean; therefore, the probability of observing an error of x from the first source and y from the second source has to be the joint probability, i.e.,

$$pdf_{\alpha,\beta}(x, y) = \frac{N}{2\pi\sigma_\alpha\sigma_\beta} e^{-N(x^2/2\sigma_\alpha^2 + y^2/2\sigma_\beta^2)} \quad (72)$$

Now for such a combination of errors the total error is just $(x + y)$. So what is the average square of the error, i.e., the variance, over all possible x and y? This is just the two dimensional averaging (i.e., integral) of $(x + y)^2$, weighted by $pdf_{\alpha,\beta}(x, y)$, i.e.,

$$\mathrm{var}(x + y) = \frac{N}{2\pi\sigma_\alpha\sigma_\beta} \int\int_{\substack{x=-\infty,\infty \\ y=-\infty,\infty}} (x + y)^2 e^{-(N/2)(x^2/\sigma_\alpha^2 + y^2/\sigma_\beta^2)} \mathrm{d}x\mathrm{d}y \quad (73)$$

We can split this into three integrals by expanding $(x + y)^2$. Thus:

$$\begin{aligned}\mathrm{var}(x+y) = &\frac{N}{2\pi\sigma_\alpha\sigma_\beta}\int\limits_{\substack{x=-\infty,\infty\\y=-\infty,\infty}}\int x^2 e^{-(N/2)(x^2/\sigma_\alpha^2+y^2/\sigma_\beta^2)}\mathrm{d}x\mathrm{d}y\\ &+\frac{N}{\pi\sigma_\alpha\sigma_\beta}\int\limits_{\substack{x=-\infty,\infty\\y=-\infty,\infty}}\int xy e^{-(N/2)(x^2/\sigma_\alpha^2+y^2/\sigma_\beta^2)}\mathrm{d}x\mathrm{d}y\\ &+\frac{N}{2\pi\sigma_\alpha\sigma_\beta}\int\limits_{\substack{x=-\infty,\infty\\y=-\infty,\infty}}\int y^2 e^{-(N/2)(x^2/\sigma_\alpha^2+y^2/\sigma_\beta^2)}\mathrm{d}x\mathrm{d}y\end{aligned} \tag{74}$$

We can rewrite the first term as:

$$\begin{aligned}&\frac{N}{2\pi\sigma_\alpha\sigma_\beta}\int\limits_{\substack{x=-\infty,\infty\\y=-\infty,\infty}}\int x^2 e^{-(N/2)(x^2/\sigma_\alpha^2+y^2/\sigma_\beta^2)}\mathrm{d}x\mathrm{d}y\\ &=\frac{N}{2\pi\sigma_\alpha\sigma_\beta}\int\limits_{x=-\infty}^{\infty}x^2 e^{-(N/2)(x^2/\sigma_\alpha^2)}\mathrm{d}x\int\limits_{y=-\infty}^{\infty}e^{-(N/2)(y^2/\sigma_\beta^2)}\mathrm{d}y\end{aligned} \tag{75}$$

Therefore, we can integrate the integral over y independently. We can do the same thing for the third term for x. This leads to:

$$\begin{aligned}\mathrm{var}(x+y) = &\sqrt{\frac{N}{2\pi\sigma_\alpha}}\int\limits_{x=-\infty,\infty}\int x^2 e^{-(N/2)(x^2/\sigma_\alpha^2)}\mathrm{d}x\\ &+\frac{N}{\pi\sigma_\alpha\sigma_\beta}\int\limits_{\substack{x=-\infty,\infty\\y=-\infty,\infty}}\int xy e^{-(N/2)(x^2/\sigma_\alpha^2+y^2/\sigma_\beta^2)}\mathrm{d}x\mathrm{d}y\\ &+\sqrt{\frac{N}{2\pi\sigma_\beta}}\int\limits_{y=-\infty,\infty}\int y^2 e^{-(N/2)(y^2/\sigma_\beta^2)}\mathrm{d}y\end{aligned} \tag{76}$$

Now, given that the mean is zero, the first term is just the integral for the variance due to x, the third term is the integral for the variance due to y, and the second term must be zero because it separates into the product of two integrals each of which must be zero as they calculate the average value of x and y, respectively, both zero. Therefore:

$$\mathrm{var}(x+y) = \mathrm{var}(x) + \mathrm{var}(y) \tag{77}$$

The astute reader will notice that we could have performed the same sequence of operations with any *pdf*, not just a Gaussian, and arrived at the same conclusion. The key steps are multiplying the individual *pdf*s together, separating the resultant integral into three integrals, two of which are the individual variances and the third must be equal to zero because we defined the mean of each *pdf* to zero. That is, this is a general result, not just one pertaining to Gaussian forms of the distribution function.

Appendix 3: Deriving the Hanley Formula for AUC Error

Recall that we have the following equation for the variance of either the actives or the inactives, where w = AUC for the former and $w = 1 - \text{AUC}$ for the later:

$$\text{Var} = \frac{w^2(1-w)}{1+w} \tag{78}$$

The assumption by Hanley is that the *pdf* for both actives and inactives follows an exponential distribution, e.g.

$$\begin{gathered} x \geq 0 \\ p_{\text{active}}(x) = \lambda_{\text{a}} e^{-\lambda_{\text{a}} x} \\ p_{\text{inactive}}(x) = \lambda_{\text{i}} e^{-\lambda_{\text{i}} x} \\ x < 0 \\ p_{\text{active}}(x) = p_{\text{inactive}}(x) = 0 \end{gathered} \tag{79}$$

Here, x is a score for either active or inactive that determines its rank (higher = better). These forms integrate from 0 to positive infinity to 1.0 as required. Since we can always rescale x by a constant and still have the same rankings, let's set the lambda for inactives to 1, i.e.,

$$\begin{aligned} p_{\text{active}}(x) &= \lambda e^{-\lambda x} \\ p_{\text{inactive}}(x) &= e^{-x} \end{aligned} \tag{80}$$

Given these probability density functions we can write down an expression for the AUC either in terms of the fraction of inactives of lower score than each active, or as the fraction of actives higher than each inactive. The math is a little cleaner if we do the latter:

$$\text{AUC} = \int_{x=0}^{x=+\infty} p_{\text{inactive}}(x) \int_{y=x}^{y=+\infty} p_{\text{inactive}}(y)\mathrm{d}y\mathrm{d}x \tag{81}$$

The first term in the integral is the density of inactives at score x and this is multiplied by the fraction of actives with a score greater than x. If we substitute the Hanley *pdf*s we get:

$$\begin{aligned} AUC &= \int_{x=0}^{x=\infty} e^{-x} \int_{y=x}^{y=\infty} \lambda e^{-\lambda y}\mathrm{d}y\mathrm{d}x \\ &= \int_{x=0}^{x=\infty} e^{-x} e^{-\lambda x}\mathrm{d}x \\ &= \frac{1}{1+\lambda} \end{aligned} \tag{82}$$

We can see that this looks correct because if lambda is greater than one the scores of the actives must fall off more quickly than the inactives and the AUC will be less than 0.5, but if it is less than one it has a longer tail of positive scores and so has an AUC greater than 0.5. Now, let's consider the variance for the inactives:

$$\mathrm{Var}_{\mathrm{inactives}} = \int_{x=0}^{x=\infty} e^{-x} \left[\int_{y=x}^{y=\infty} \lambda e^{-\lambda y} \mathrm{d}y \right]^2 \mathrm{d}x - \mathrm{AUC}^2 \tag{83}$$

This is just the equivalent of $<p^2> - <p>^2$ we normally see for a variance but we are integrating over the *pdf* for inactives. Expanding and solving the integral we get:

$$\begin{aligned} \mathrm{Var}_{\mathrm{inactives}} &= \int_{x=0}^{x=\infty} e^{-x} e^{-2\lambda x} \mathrm{d}x - \mathrm{AUC}^2 \\ &= \frac{1}{1+2\lambda} - \mathrm{AUC}^2 \end{aligned} \tag{84}$$

Now, the nice thing about the Hanley choice is that we can substitute for lambda from the AUC, i.e.,

$$\begin{aligned} \mathrm{AUC} &= \frac{1}{1+\lambda} \\ \lambda &= \frac{1-\mathrm{AUC}}{\mathrm{AUC}} \\ \mathrm{Var}_{\mathrm{inactives}} &= \frac{\mathrm{AUC}}{2-\mathrm{AUC}} - \mathrm{AUC}^2 \\ &= \frac{\mathrm{AUC}(1-\mathrm{AUC})^2}{2-\mathrm{AUC}} \end{aligned} \tag{85}$$

Setting $w = 1 - \mathrm{AUC}$, we get:

$$\mathrm{Var}_{\mathrm{inactives}} = \frac{w^2(1-w)}{1+w} \tag{86}$$

And the result required is obtained. A further nice thing about the Hanley *pdf* is that we can get a simple expression for the ROC curve. If we want to know what fraction, f, of inactives or actives have a score greater than z we have:

$$\begin{aligned} f_{\mathrm{inactive}}(z) &= \int_{y=z}^{y=\infty} e^{-y} \mathrm{d}y \\ &= e^{-z} \\ f_{\mathrm{active}}(z) &= e^{-\lambda z} \end{aligned} \tag{87}$$

But $(f(z)_{\text{inactive}}, f(z)_{\text{active}})$ are the points on the ROC curve, parameterized by z. Therefore, to express the one in terms of the other we simply have:

$$\begin{aligned} x &= e^{-z} \\ y &= e^{-\lambda z} \\ \therefore y &= x^{\lambda} \\ y &= x^{\frac{1-\text{AUC}}{\text{AUC}}} \end{aligned} \tag{88}$$

This is the form of the Hanley ROC curve for a given AUC value. It can be a pretty good fit to real data!

References

1. Loredo, T. J., From Laplace to Supernova SN 1987A: Bayesian inference in Astrophysics. Maximum Entropy and Bayesian Methods. P. F. Fougere (ed). Kluwer Academic, Netherlands: 1990, 81–142.
2. Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. 3rd ed; Cambridge University Press, New York: 2007.
3. Wainer, H., The most dangerous equation: Ignorance of how sample size affects statistical variation has created havoc for nearly a millennium. *Am. Sci.* 2007, 248–256.
4. Stigler, S. M., Statistics and the question of standards. *J. Res. Natl. Inst. Stand. Technol.* 1996, **101**, 779–789.
5. Student, The probably error of a mean. *Biometrika* 1908, **6**, 1–25.
6. DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L., Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988, **44**, 837–845.
7. Cortes, C.; Mohri, M., Confidence intervals for the area under the ROC curve. *Adv. Neural. Inf. Process. Syst.* 2004, **17**, 305–312.
8. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking sets for molecular docking. *J. Med. Chem.* 2006, **49**, 6789–6801.
9. Bayly, C. I.; Truchon, J.F., Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.*, 2007, **47**, 488–508.
10. Jain, A. N., Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* 2007, **21**, 281–306.
11. Skillman, A. G.; Nicholls, A., *SAMPL2: Statistical Analysis of the Modeling of Proteins and Ligands*: 2008.
12. Scargle, J. D., Publication bias: The "File-Drawer" problem in scientific inference. *J. Sci. Explor.* 2000, **14**, 91–106.
13. Ziliak, S. T.; McCloskey, D. N., *The Cult of Statistical Significance*. The University of Michigan Press, USA: 2007.
14. Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S., A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 2006, **49**, 5912–5931.
15. Enyedy, I. J.; Egan, W. J., Can we use docking and scoring for hit-to-lead optimization? *J. Comput. Aided Mol. Des.* 2008, **22**, 161–168.
16. Rerks-Ngarm, S.; Pitisuttithum, P.; Nitayaphan, S.; Kaewkungwal, J.; Chiu, J.; Paris, R.; Premsri, N.; Namwat, C.; de Souza, M.; Adams, E.; Benenson, M.; Gurunathan, S.; Tartaglia, J.; McNeil, J. G.; Francis, D. P.; Stablein, D.; Birx, D. L.; Chunsuttiwat, S.; Khamboonruang, C.; Thongcharoen, P.; Robb, M. L.; Michael, N. L.; Kunasol, P.; Kim, J. H., Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N. Engl. J. Med.* 2009, **361**, 2209–2220.
17. Welch, B. L., The generalization of "student's" problem when several different population variances are involved. *Biometrika* 1946, **34**, 28–35.
18. Satterhwaite, F. E., An approximate distribution of estimates of variance components. *Biometrics Bull.* 1947, **2**, 110–114.
19. Glantz, S. A., How to detect, correct, and prevent errors in the medical literature. *Circulation* 1980, **61**, 1–7.

20. Snedecor, G. W.; Cochran, W. G., *Statistical Methods*. 8th ed.; Blackwell Publishing, Malden, MA: 1989.
21. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K., Gaussian docking functions. *Biopolymers* 2003, **68**, 76–90.
22. Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A., A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* 2005, **48**, 1489–1495.
23. Glantz, S. A., *Primer of Biostatistics*. 5th ed.; McGraw-Hill, New York: 2002.
24. Kanji, G. K., *100 Statistical Tests*. 3 rd ed.; Sage Publications, London: 2006.
25. Bulmer, M. G., *Principles of Statistics*. Dover, USA: 1979.
26. Keeping, E. S., *Introduction to Statistical Inference*. Dover, USA: 1995.
27. van Belle, G., *Statistical Rules of Thumb*. Wiley, New York: 2002.
28. Pepe, M. S., *The Statistical Evaluation of Medical Tests for Classifaction and Prediction*. Oxford University Press: 2004.
29. Good, P. I.; Hardin, J. W., *Common Errors in Statistics (and How to Avoid Them)*. 2nd ed.; Wiley-InterScience, New Jersey: 2006.
30. Moye, L. A., *Statistical Reasoning in Medicine*. 2nd ed.; Springer, New York: 2006.
31. Silvia, D. S., *Data Analysis: A Bayesian Tutorial*. Oxford Science Publications: 1996.
32. Marin, J. -M.; Robert, C. P., *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York: 2007.
33. Carlin, B. P.; Loius, T. A., *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd ed.; Chapman & Hall/CRC, Boca Raton, FL: 2000.
34. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G., Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci*. 2002, **42**, 1273–1280.
35. Vidal, D.; Thormann, M.; Pons, M., LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Complement. Integr. Med*. 2005, **45**, 386–393.

INDEX

A

Activity cliffs 19, 44, 77–78, 80, 82, 84, 85, 93, 103–105, 107, 109, 112, 113, 115, 116, 119, 120, 122, 130, 222, 223, 268
Activity landscapes 18, 19, 44, 77–78, 83–85, 93, 101–116
ADME/ADMET 4, 5, 19, 178, 200, 215, 219, 389, 390, 393–395, 408, 409, 413, 420, 424
Adverse drug reaction 187, 191–193
Alignment .. 3, 4, 17, 21, 22, 40, 43, 50, 61–63, 66–70, 92, 263, 274–281, 303–305, 326, 329, 363, 364, 374, 421. *See also* Superposition/overlay
ANOVA 534, 566–571
Applicability domain (AD) 18, 215, 225–230, 284, 477, 497
Area under the curve (AUC) 347, 533, 542, 543
Artificial neural network (ANN) 183, 186, 193, 344, 348
Atom environments 216
Atom pairs 50, 64, 216–221, 223, 392, 395
Augmented atoms 216, 230
Autodock 11

B

Bayesian
  classifier 149, 150, 177, 179, 184–187, 189, 190, 193, 223, 281, 446
  statistics 532, 574
Bayes' theorem 162, 176, 180, 193
BCI 220, 221
BCUTs 45, 57, 58, 392, 398, 399, 401, 420
Binary QSAR 180, 189
Bioisosterism 276
Biological networks 478, 484
BioPath 20, 326
BioPredsi 348, 349, 351, 354
BioXyce 470, 478, 479, 482–484
Bit string 122, 140, 159, 214, 221, 402, 492
BLASTn 343
Blood–brain barrier 224, 225, 289, 394, 446
BOMB 13
BREED 303, 304

C

CAMEO 3
CASREACT 2
CAS Registry 5, 9, 10, 246
Catalyst 3, 4, 21, 22, 264, 270, 282, 286–288
Central Limit Theorem (CLT) 534–542, 558, 563, 565, 573, 576
ChemBank 375, 468
ChemBridge structure database 25
Chemical genetics 468, 503
Chemical information 1, 2, 140, 197, 234, 281, 327, 468, 470
Chemical library 5
Chemically advanced template search (CATS) 220, 281, 317
Chemical space
  cell-based 41
  coordinate-based 41, 44, 73, 74, 93
  coordinate-free 44, 73, 74, 93
Chemical structure representation 9–11
ChemInform RX 330
Chemogenomics 138, 185, 213, 448, 467, 469, 503, 517
Chemotype 286–288, 292, 301, 306, 320, 406, 419, 420, 422, 440, 441, 443, 445, 465
ChemReader 25
ChemSpider 7, 9, 10, 470
Circular fingerprints 178, 181, 182, 184–186, 190, 216, 223
Classical sets 42, 43, 45, 91, 93–95
ClassPharmer 17, 23, 409, 411
Clique detection 17, 21, 280
CLOGP 224
Clustering 5, 17, 24, 70, 92, 115, 121, 123, 125, 137, 143, 146, 205–206, 226, 231, 249, 253–254, 257–259, 331, 336, 338, 395–398, 400, 416, 445, 447–450, 571
COLIBREE 319
CombiDOCK 412, 413, 418
CombiGLIDE 394, 412–414, 418, 422
Combinatorial library 16, 230, 250, 388, 392, 396, 399, 400, 402–406, 409, 412, 415, 419, 440
CoMFA/CoMSIA 375
Comparative molecular field analysis (CoMFA) ... 4, 17, 23, 46, 269, 375
Compound
  activity classes 120, 121, 123, 125, 165, 166, 168, 169
  database 372, 468
  recall 164–167
CONCERTS 305
Condensed reaction graph (CRG) 231–233, 329
CONFIRM 319

Conformational
flexibility 41, 43, 62, 67–69, 92, 319
sampling.....13, 27, 67, 263, 266, 270, 279, 280, 282, 283, 285, 290, 291, 402
search 22, 62, 68, 70, 424
space.....4, 22, 69
Conformer.....17, 43, 69–70, 92, 263, 264, 269–272, 274, 275, 277, 279, 280, 282–284, 291, 313, 421
Consensus activity cliffs 19, 85
Cross validation..... 17, 18, 113, 187, 226–229, 233, 335, 347, 534
Crystal structure.....3, 12, 22, 27, 188, 272, 273, 275, 314, 359–382, 391, 406, 414, 422, 474. *See also* X-ray crystallography

D

Data
fusion 16, 42, 83, 147–149, 160, 161, 221, 447
reduction 23–24
visualization 23–24, 510
Data base searching
2D 2, 135, 146
3D 3, 6
Daylight fingerprint 140, 169, 201, 203, 205, 210, 220, 334, 392
Decision tree..... 20, 206, 344, 347
De novo design 3, 13–14, 26, 299–303, 306, 308–320, 423
DEREK..... 9, 219, 223
Descriptors 4, 41, 102, 137, 181, 200, 213–235, 246, 268, 326, 345, 388, 437, 467, 490, 513, 519
Design of Genuine Structures (DOGS) ... 309–311, 467
Desolvation.....13
DISCO..... 3, 22
Dissimilarity..... 5, 40, 42, 44, 66, 68, 69, 71, 75, 77, 92, 223, 268, 269
DiverseSolutions 394, 398–401, 419, 420
DOCK 6, 11, 189, 288, 372, 381, 412, 423, 474
DOGS. *See* Design of Genuine Structures
Dose-response 24, 106–110
Drugability 15
Drug design
fragment-based.....14–16, 26
ligand-based.....3, 4
structure-based 3, 14, 15, 26, 193
Drug discovery.....2, 5, 9, 13, 20, 21, 119, 131, 135, 136, 146, 149, 179, 180, 184, 188, 193, 246, 266, 273, 299, 310, 314, 318, 320, 355, 361, 362, 389, 411, 413, 415, 425, 435, 436, 446, 460, 461, 464, 490, 501, 503, 531, 548
Druglikeness..... 4–6, 14
DUD error propagation 556, 559

E

ECFP4 492
Eigenvalue 76, 77
Entropy.....13, 27, 82, 221, 222, 495–497
Enzyme commission (EC).....327, 331, 333, 335–338
EON 422, 423
EROS..... 3
E-state descriptors.....395
Euclidean distance.....43, 50, 51, 59, 77, 91, 109, 143, 335, 449

F

FBDD. *See* Fragment-based drug design
Feature
tree 75, 198, 209, 210, 313
vector 50–52, 55–61, 90, 91, 177, 193, 333
Fields
electrostatic..... 139, 269, 423
lipophilic 4, 61
steric..... 64, 313
Fingerprint
2D 16, 135, 136, 138, 147, 148, 181, 184, 201, 205, 206, 423, 523, 528
3D 279, 282, 402, 403, 406–408
FLAP.....406, 422
Flexibility
ligand 12, 270, 275
protein 13, 271–272
Flexophore..... 269. *See also* Pharmacophore
FlexX..... 11, 23, 189, 312, 413, 418
FLOG 11
FLUX.....307, 317, 318
Focused compound library..... 442–444, 451
FOG.....306, 307
Force field 11, 262, 263, 269, 270, 274, 283, 284, 305–306, 556
Formal concept analysis 504, 510–513
Fourier transform..... 68, 92
Fragment ...14, 45, 133, 159, 179, 197, 214, 254, 270, 301, 329, 391, 442, 492, 511, 548
Fragment-based drug design (FBDD) 14–16, 27, 415, 416, 419
Fragment descriptor.....213–235
Frameworks .....9, 24, 27, 42, 64, 82, 111, 125, 160, 161, 180, 183, 205, 206, 218, 225, 249, 277, 286, 409, 410, 478, 480, 510, 515. *See also* Scaffolds
FRED.....11, 288, 480, 570–572
Free energy 16, 26, 262, 269, 270, 272, 345
Free–Wilson analysis 214
Fuzzy pharmacophore triplets (FPT) 217, 221, 223

G

GALAHAD. *See* Genetic Algorithm with Linear Assignment for Hypermolecule Alignment of Datasets
GASP. *See* Genetic Algorithm Superimposition Program
Gaussian ...20, 43, 46, 61–64, 180, 181, 192, 274, 275, 421, 423, 444, 523, 525, 527, 528, 533, 534, 541, 554, 558, 562, 563, 566, 569, 575, 577, 5360–539
Genetic algorithm 4, 11, 59, 111, 112, 204, 225
Genetic Algorithm Superimposition Program (GASP)... 3, 4, 22
Genetic Algorithm with Linear Assignment for Hypermolecule Alignment of Datasets (GALAHAD)..... 21, 22, 288
GLIDA. *See* GPCR–Ligand Database
GLIDE..... 11, 189, 286, 372, 375, 377, 422
GOLD 11, 288, 373, 474
GPCR–Ligand Database (GLIDA).....469
G protein coupled receptor (GPCR)..... 187, 188, 289, 359–382, 389–391, 400, 401, 406–411, 441, 464, 469, 473, 497, 499, 501
Graph variables
  Boolean.....42
  categorical.....42, 50, 56, 177
  non-negative integer 56
  real vector 42

H

Hamming distance 50, 51, 59, 91
Hashing 50, 56, 140
High-throughput screening (HTS) 135, 180, 184, 186, 205, 207, 246, 249, 388, 460, 490
Hill equation 108, 109
HipHop 3, 264
Homology modeling 376, 391, 442
Hydrogen bond
  acceptor.....4, 14, 183, 211, 217, 218, 220, 265, 267, 318, 371, 399, 416, 421
  donor 4, 14, 183, 202, 207, 211, 217–220, 263, 266, 399, 402, 416, 421
Hydrogen-suppressed graph 47
Hyperplane 518–522

I

ICM 11, 373
IGOR.....3
InChI. *See* International Chemical Identifier
Inductive logic programming (ILP) 223, 280, 281
In silico target profiling 497–499
International Chemical Identifier (InChI) ... 10, 11, 247
ISIDA.....215, 216, 224, 225, 230, 233, 234
ISIS 54, 349, 409

K

KEGG 327, 328, 335, 336, 468–470
Kernels 521, 523, 528
Kinases .. 185, 187, 289, 389, 390, 401, 405, 409, 411, 441, 497
k-nearest neighbors (KNN).....450
KNIME.....7, 9
Kohonen 330–333, 336, 337
Kohonen neural network.... 330. *See also* Self-organizing map
Kullback–Leibler divergence 161, 163–169, 171

L

Lagrange multipliers 520
LAMDA.....21
Laplacian correction.....117, 178, 184
Leadlikeness..... 6, 14
Lead optimization .. 119, 120, 131, 178, 389–391, 393, 394, 415, 417, 436, 450
LHASA 2, 8, 9
Library design
  cell-based 397–401
  combinatorial 16, 387–425
  multi-objective 5
  parallel..... 387–425
  target-focuses 403–406
LigandScout 21, 286
Ligand-target interactions 219
LigBuilder.....316, 317, 320
Likelihood estimate..... 19, 161, 283, 490
Linear discriminant analysis (LDA).....190, 450
Linear regression 111, 112, 161, 165–168, 230, 281, 308, 344, 349, 375
LIQUID 286
LUDI 13, 189

M

MACCS structural keys 121, 169, 492, 519
Machine learning.....3, 20, 25, 26, 103, 111, 115, 116, 134, 141, 145, 147, 149, 150, 160, 178, 181, 183–185, 214
Markov chain 306
Markush structures 2, 26, 197–200, 211
Matched molecular pairs.....102
Maximum common
  subgraph 16, 48, 205
  substructure....16, 17, 47, 48, 50, 62, 90, 206, 410, 447
MCSS.....303
MED-hybridise.....303, 304
Metabolite prediction 20–21
Metropolis Monte Carlo.....301
MM-GBSA 375, 376

MODDE 396
Mode of action 138, 489
MOE. *See* Molecular Operating Environment
MolconnZ 225, 395
Molecular
  diversity 71, 135, 137, 393
  dynamics 11, 69, 271, 290, 305, 370
  frameworks .... 24, 218 (*see also* Scaffolds)
  similarity 4, 16–17, 39–95, 121–122, 136, 221, 439, 507
  surface 139, 182
Molecular Operating Environment (MOE) 21, 22, 122, 166, 189, 217, 257, 287, 288, 353, 354, 377
MOLMAP 334–338
Molprint 2D 181, 182, 184
Morgan algorithm 9, 10
MultiCase 219
Multi-objective genetic algorithm (MOGA) 22
Multi-objective optimization 22
Multiple linear regression 230
Multi-task learning (MTL) 227, 228

N

Neighborhood behavior 102
Network 7, 26, 84, 105, 106, 109, 110, 113–115, 121, 125, 126, 128–131, 178, 179, 183, 186, 193, 227, 230, 246, 330, 349, 374, 401, 459, 460, 463, 468, 470, 474, 475, 478, 480, 482–484, 499
NMR 14, 15, 272, 417–419
Non-linear mapping 75, 76
Normalization 123–125
Nuclear hormone receptors 389, 400, 409, 411, 441

O

Off-target effects 446, 448, 450
Ontology 390, 411, 448, 473
Open systems 8–9
Optical Structure Recognition Software (OSRA) 25
Orthogonality 88, 89
OSRA. *See* Optical Structure Recognition Software

P

P-value 114
Pareto ranking 22, 254, 277
Partial equalization of orbital electronegativity (PEOE) 330,335
Partitioning 5, 185, 189, 199, 206, 209, 253–254, 282, 446, 462
PEOE. *See* Partial equalization of orbital electronegativity
PETRA 338
Pharmacodynamics 219, 465
Pharmacokinetics 214, 219, 227, 265, 353, 489
Pharmacophore 3, 4, 6, 11, 13, 21, 22, 52, 111, 115, 116, 134, 159, 166, 217, 221–223, 250, 261–292, 395, 401–409, 420, 422, 423, 442, 475
Pharmacophoric fragment (PHRAG) 492, 493, 497
PharmID 21
PharmPrint™ 403, 405
Phase 21, 22, 289, 387, 414, 436, 471, 525
PHRAG. *See* Pharmacophoric fragment
Pipeline Pilot 7, 17, 254, 409
Polypharmacology 460, 461
Population analysis 41
Precision 145, 208, 473, 570–572
Preference ranking 518, 519, 521–526, 528, 529
Principal component analysis (PCA) 41, 58, 72, 76, 77, 188, 353, 395, 403
Privileged substructures ... 389, 393, 395, 407, 409–411
Probability distribution 533, 536
Protein Data Bank (PDB) 3, 138, 191, 303, 469, 470, 480
Protein–ligand docking 6, 11, 13
PubChem 8, 9, 246, 250, 251, 304, 460, 468, 470
PubMed 465

Q

QSAR/QSPR 1–4, 16–20, 22, 26, 56, 78, 80, 83, 93, 101, 103, 109, 111, 136–138, 142, 146, 180, 186, 189, 214, 215, 218, 219, 222–233, 235, 264, 268, 275, 279–281, 288, 329, 335, 343, 344, 353, 354, 388, 390, 394, 401, 403, 405, 408, 438, 445, 463, 475–478, 489, 490, 497, 498

R

R-group 23, 24, 102, 396, 412, 413, 420, 423
Random forest 26, 278, 351
RASSE 302
REACCS 2
Reaction
  center 14, 309, 326, 328–331, 333, 335, 338
  classification 327, 328, 330
  descriptors 328, 329, 333
  retrieval 2, 325
  similarity search 232
Reaction-MQL 309
RECAP. *See* Retrosynthetic combinatorial analysis procedure
Receiver operator characteristic (ROC) 13, 145, 229, 347, 349, 533, 542, 543, 545, 548, 550, 551, 553–555, 573, 574, 579, 580
Recore 319

Recursive partitioning ..... 185, 189, 206, 446. *See also* Decision tree
Reduced graphs 197–211, 218, 249, 309
Reference
molecules 59, 66–68, 160–162, 164, 166, 167, 313, 528
structure 16, 134–136, 138, 141–143, 145–148, 221
Retrosynthetic combinatorial analysis procedure (RECAP) 215, 307, 308, 313
Retrosynthetic methods 14, 215, 307–308
ROCS 13, 288, 422, 423, 571, 572
Route designer 308
Route mean square deviation (RMSD) 13, 277, 278, 367, 370, 376–378
Rule-of-five 70, 389, 395, 416

S

SAR. *See* Structure–activity relationships
Scaffolds
hopping 80–82, 141, 146, 186, 200, 205, 209–211, 246, 265, 266, 269, 281, 286, 287, 292, 319, 391, 395, 400, 420, 421, 423–425, 441
tree 23, 24, 245–259, 447
Scaffold hopping (SHOP) 421–423
SciFinder 24, 465
Scoring ... 11–14, 22, 23, 116, 120–122, 125, 131, 147, 149, 166, 188–190, 221, 222, 264, 265, 269, 272–275, 278, 284, 288, 291, 300, 301, 306, 308, 312–314, 316–319, 343, 346, 348, 372–376, 381, 394, 412–414, 423, 424, 474, 490, 558
SEA. *See* Similarity Ensemble Approach
Selectivity
cliff 119–131
searching ... 504, 507, 510, 518, 521, 523, 525, 528
Self-organizing map 23, 330
Similog keys 217
Shannon entropy 82, 221, 222, 495–497
Shannon entropy descriptors (SHED) 492, 495–499
Shape 11, 13, 40, 41, 61, 75, 139, 245, 274, 282, 288, 313, 372, 377, 391, 406, 420–423, 531
complementarity 262
indices 41, 58
SHED. *See* Shannon entropy descriptors
SHOP. *See* Scaffold hopping
Short interfering RNA (siRNA) 341–356, 437, 444, 448
siDESIGN 352
Similarity
coefficient 16, 42, 43, 51–53, 56, 60, 61, 89, 90, 104, 134, 135, 141–143, 148, 160, 188
indices 42, 47, 49, 52–55, 58–61, 65–67, 87, 88, 90, 91
measure 39–95, 104, 134, 135, 137–139, 144–148, 151, 161, 204, 209, 220, 221, 258, 301, 313, 328, 447, 495
searching 2, 16, 52–55, 91, 93, 133–151, 159–171, 200–201, 204, 209, 210, 214–218, 220–223, 231, 232, 281, 403, 445, 448, 517
Similarity Ensemble Approach (SEA) 475, 499, 501
Similarity property principle (SPP) 119, 120, 160, 507
Simulated annealing 11, 111, 311, 401, 404
siRNA. *See* Short interfering RNA
SitePrint 406
SkelGen 319, 415
SMARTS 202, 207, 208, 211
SMILES 10, 203, 247, 467
SMIRKS 309
Solubility 57, 102, 219, 224, 226, 234, 301, 394, 416, 442, 467, 470, 472, 490
SOSA 451
Spiral View 24
SPP. *See* Similarity property principle
SPROUT 13, 314, 315, 320
Standard deviation 71, 535, 537, 542, 562, 563
Statistical
analysis 140
experimental design 388, 395, 396, 422
Statistics 87, 113, 161, 187, 306–307, 477, 531–573
Stereo center 308
Structural interaction fingerprint (SIFt) 23
Structure–activity index (SARI) .. 19, 84, 104, 106–108, 113, 115, 120, 122
Structure–activity landscape index (SALI) 19, 84, 104–106, 109–116
Structure–activity relationships (SAR)
continuity 107, 122
discontinuity 103, 107, 115, 116, 122–124, 126, 128–131
heterogeneity 19, 106, 120, 126, 507
maps 23, 24
multi-target 120
by NMR 14, 417
Structure–activity similarity (SAS) 44, 78–83
Structure–selectivity relationship (SSR) 120–126, 128–131, 507, 510
Student *t*-test 534, 539, 561, 566–571
Subgraph isomorphism 414
SubScape 24
Substituent 40, 129, 199, 230, 309, 312, 371, 397, 413
Substructure
analysis 16–17
searching 2, 135, 136, 139–141, 151, 214

Superposition/overlay 274–279, 286, 362
Support vector machine (SVM) 20, 150, 178, 183, 185, 193, 217, 223, 230, 233, 234, 335, 344, 347, 350, 354, 446, 450, 510, 517–529
Surflex 373
Sybyl 287, 288, 334, 410, 493, 494
SYLVIA 307, 308
SYNOPSYS 308, 309
Synthesis design 2
Synthetic feasibility 14, 106, 308, 425
Systematic Generation of Metabolites (SyGMa) 21
Systems biology 355, 459–485

T

Tanimoto coefficient (T_c) 122, 142, 143, 148, 151, 160, 164, 169, 201, 232, 406, 509
Target
  class combinatorial libraries 390
  family library 389, 390, 392, 393, 401, 403, 406–409, 411
  prediction 190
Text mining 25, 26
TOPAS 307
TopKat 219
Topological torsions 50, 218, 223
Topology ... 12, 59, 115, 126, 139, 198, 203, 211, 215, 246, 304, 337, 360, 363, 365, 367, 368, 380, 495
Topomer 17, 313
Turbo similarity searching 16, 148
Tversky coefficient 143

U

UNITY 2D 220, 221

V

van der Waals 4, 262
Variance 58, 62, 72, 123, 225, 353, 534–537, 541–545, 547, 553–560, 564–571, 574, 576–579
Virtual screening
  ligand-based 13, 18, 23, 159, 160, 214, 235
  pharmacophore-based 261–292
  structure-based 23, 214, 474

W

WODCA 2
World of Molecular Bioactivity (WOMBAT) 12, 13, 188, 190, 191, 209, 246, 249, 252, 257, 460, 467, 474, 497
World Wide Web 7

X

X-ray crystallography 14, 15, 267, 416–418, 480

Z

Z-score 124
ZINC 9, 12, 166, 167, 288, 372, 468, 477