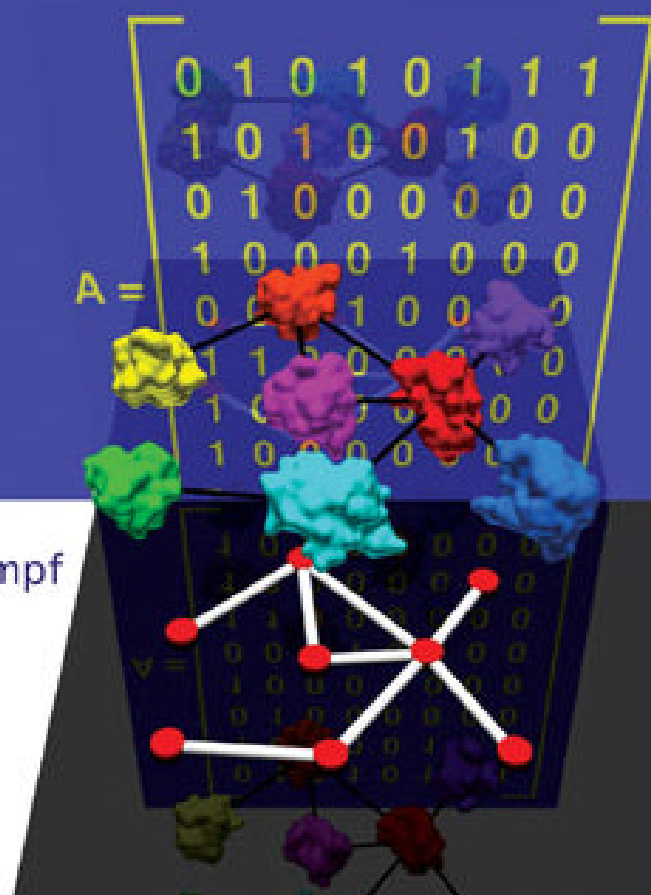


Michael P H Stumpf
Carsten Wiuf
editors



Statistical and Evolutionary Analysis of **Biological Networks**

Imperial College Press

Statistical and Evolutionary Analysis of Biological Networks

This page intentionally left blank

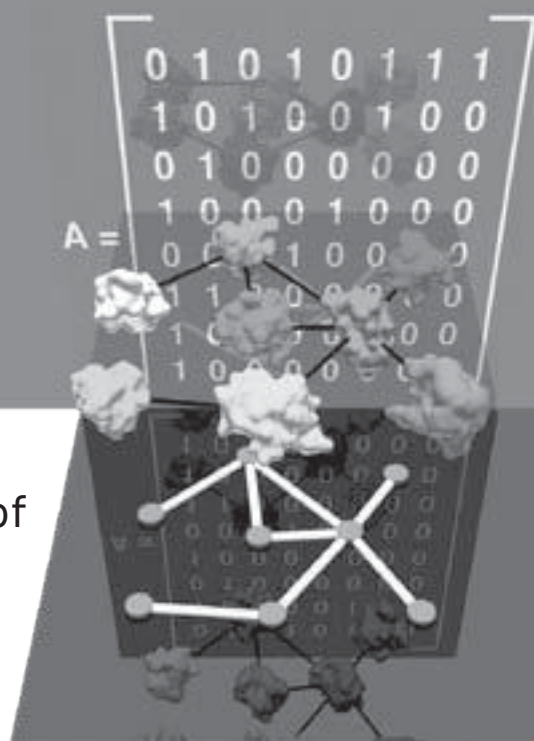
editors

Michael P H Stumpf

Imperial College London, UK

Carsten Wiuf

Aarhus University, Denmark



Statistical and Evolutionary Analysis of Biological Networks

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

STATISTICAL AND EVOLUTIONARY ANALYSIS OF BIOLOGICAL NETWORKS

Copyright © 2010 by Imperial College Press

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the Publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

ISBN-13 978-1-84816-433-8
ISBN-10 1-84816-433-5

Printed in Singapore.

Preface

In recent years many new data types and settings have become available through new large-scale and high-throughput technologies, but also through initiatives that seek to collect biological and epidemiological data in society at large. These data types provide new perspectives on the organisation, complexity, functionality and dynamics of biological entities and potentially offer a deeper insight into what constitutes a cell or organism, and how cells, organisms and species are related through common origin, evolution and development.

However, the new data types are by themselves exceedingly complex and barely understandable without further processing or analysis. Many of the new data types, such as transcriptomic, metabolomic and protein interaction data, have provided means to define corresponding new ‘omes’ – for example, the transcriptome, metabolome and interactome – that not only reflect the data type and technology, but also structure the functionality and organisation of the organism conceptually. In relation to this, mathematical theory, in particular network theory, has been essential and proven an indispensable tool for understanding and interpreting data.

A link in a network or graph represents an interaction between two entities; the interaction could represent direct physical contact, e.g. the binding of two molecules to each other, that the presence of one molecule stimulates the presence of another molecule, or a path through which a disease can spread. We are becoming accustomed to talking about ‘biological networks’ or ‘biological network data’ and by this we mean the relevant biological data structured by a network interpretation. The biological network data is not the ‘raw’ biological data, but the data imposed onto a network.

Apart from their apparent usability for visualisation of highly interdependent data, networks allow stringent mathematical and statistical analysis. Network or graph theory goes back to Leonard Euler with his famous example of the seven bridges of Königsberg and has since proven its usefulness in numerous connections and a diverse set of different academic disciplines. A large body of graph theory exists and evolutionary, statistical and computational methods have over the last 50 years been developed to facilitate analysis of network data. Some of these developments have already been incorporated into analysis of biological network data, while at the same time new methods have been developed and applied to data.

These methods and their application to biological questions and issues are the

subject of this book. It reviews and explores statistical, mathematical and evolutionary theory and tools for understanding biological networks. It is divided into comprehensive and self-contained chapters that each focuses on an important biological network type, explains concepts and theory and illustrates how concepts and theory can be used to obtain insight into biologically relevant processes and questions. Keywords are complexity, organisation and dynamics of networks – how they come about, can be detected and measured, and how they are influenced by network evolution and functionality. The book has chapters on metabolic, transcriptomic, protein interaction and epidemiological networks, as well as chapters that deal with theoretical and conceptual material.

The authors in this volume have all contributed substantially to the discipline of network biology and we are grateful for their contributions and their patience with the editors. This is now a field which is beginning to reach maturity, and which has shaped the gestation of this volume. We hope that new investigators to this field will find the chapters in this book a useful introduction to the quantitative and evolutionary biological analysis of networks.

Contents

<i>Preface</i>	v
1. A Network Analysis Primer	1
MICHAEL P.H. STUMPF AND CARSTEN WIUF	
2. Evolutionary Analysis of Protein Interaction Networks	17
CARSTEN WIUF AND OLIVER RATMANN	
3. Motifs in Biological Networks	45
FALK SCHREIBER AND HENNING SCHWÖBBERMEYER	
4. Bayesian Analysis of Biological Networks: Clusters, Motifs, Cross-Species Correlations	65
JOHANNES BERG AND MICHAEL LÄSSIG	
5. Network Concepts and Epidemiological Models	85
ROWLAND R. KAO AND ISTVAN Z. KISS	
6. Evolutionary Origin and Consequences of Design Properties of Metabolic Networks	113
THOMAS PFEIFFER AND SEBASTIAN BONHOEFFER	
7. Protein Interactions from an Evolutionary Perspective	127
FLORENCIO PAZOS AND ALFOSO VALENCIA	

8. Statistical Null Models for Biological Network Analysis	145
WILLIAM P. KELLY, THOMAS THORNE AND MICHAEL P.H. STUMPF	
<i>Index</i>	167

Chapter 1

A Network Analysis Primer

MICHAEL P.H. STUMPF¹ AND CARSTEN WIUF²

¹*Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College
London*

²*Bioinformatics Research Center, Aarhus University
m.stumpf@imperial.ac.uk, wiuf@birc.au.dk*

Graph methods form a cornerstone of modern systems biology. In this chapter we review the fundamental apparatus of statistical descriptors and measures of graph properties. There is no single meaningful statistic that can describe all aspects of a network and we present a range of different measures that, when combined and critically evaluated, allow us to gain non-trivial insights into the architecture of complex networks in biology.

1.1. Introduction

Following the enormous advances in functional genomics and molecular biology, it is now possible to at least contemplate studying cellular processes at the level of a whole cell, rather than in isolation. Molecular networks, such as protein interaction,^{1–3} metabolic⁴ and gene regulation networks,^{5,6} aim to capture such sets of biological processes in a single and coherent framework. In reality, of course, these different networks are intricately connected and interwoven inside a cell: protein products will interact with each other, regulate the expression of genes as well as digesting nutrients and catalysing basic biochemical reactions in a cell’s metabolism. We are still a long way away from being able to consolidate these different networks into a realistic *in silico* organism.

The analysis and interpretation of present network data is, however, already challenging enough. Since the late 1990s, research has been aided considerably by the work of a host of physicists (see Refs. 7–10 for mainly physics-oriented reviews). While the models proposed have, despite their elegant simplicity, been able to explain certain aspects of complex biological networks, they increasingly reach the limit of their usefulness given the amount of data becoming available. New models, based on sound statistical principles and informed by bioinformatics, are now slowly taking their place. These networks, especially their union, form the scaffold for further systems biology investigations, and their understanding will

crucially underlie the success of the fledgling discipline of synthetic biology.

One of the central problems in the analysis of the detailed data we are confronted with now is to understand the intricate interplay between the functioning of these networks on the one hand, and their evolution on the other. While evolution clearly will not give rise to biological systems that fail spectacularly, recent research has shown that not everything found in nature has necessarily been honed by natural selection. There is indeed, as argued forcefully by Michael Lynch, a perfectly plausible explanation for any feature of biological networks in terms of a neutral evolutionary theory.

A generic problem of evolutionary analyses is, however, that evolutionary processes are highly stochastic and historically contingent. Therefore the variability inherent in evolutionary dynamics frequently masks the average behaviour and as a result, evolutionary biology has been intimately tied to statistical inference ever since it started to become a quantitative rather than a merely descriptive science. Hence the two-fold scope of this book, which puts roughly equal weight on evolutionary and statistical issues surrounding network evolution. Our aim is to present a selection of views related to how we can understand and analyse networks and their evolution¹¹ in a statistically sound manner.

1.2. Types of Biological Networks

At the molecular level we can distinguish very coarsely between three types of molecular networks.

Metabolic networks aim to describe the basic biochemistry inside a cell. Biologically important reactions have been described in terms of reaction pathways and metabolic networks are systematic collections of such biochemical data.

Transcriptional networks consist of genes where a directed edge is added between two genes if one regulates the transcription of the other gene.

Protein interaction networks in which an undirected edge is drawn between each pair of proteins where there is evidence of a physical or biochemical interaction.

Making these distinctions and simplifications must necessarily neglect details of the biological processes.¹² In reality these networks will be highly and intricately interconnected and factorising them into distinct networks will ultimately underestimate the biological complexity. These molecular networks are supplemented by physiological networks (such as the arterial and neuronal networks in higher organisms), which are not covered in this volume. Moreover, at the level of the population these networks are complemented by a higher level of networks which include food webs, ecological and epidemiological interaction and contact networks,^{13,14} and ultimately for humans, social networks.¹⁵ While we do not believe it is appropriate to push analogies which frequently do not hold up to closer scrutiny the mathematical

formalism and the statistical problems are frequently transferable. At a more ambitious level we may in fact need to include ecological interactions in order to understand the evolution and function of networks at the molecular level. This is, for example, likely to be the case when we compare different bacterial organisms, where levels of pathogenicity as well as ecological factors and type of metabolism (aerobic or anaerobic) may help to understand differences in network organisation.

1.3. A Primer on Networks

1.3.1. Mathematical descriptions of networks

Here we are primarily concerned with purely static interactions. That is, we consider the network fixed. Any changes the network might experience over time, e.g. over the life time of the organism or over evolutionary time scales, are not taken into account.

A *graph* G is the combination of a non-empty set of N nodes, $\mathcal{V}$, and a (generally but not necessarily non-empty) set of M edges, $\mathcal{E}$. In graph theory, nodes are often also called *vertices* and edges *arches*. Each edge $e_s \in \mathcal{E}$ with $1 \leq s \leq M$ is in turn associated with two nodes $v_i, v_j \in \mathcal{V}$ and we write

$$e_s = (v_i, v_j) \quad \text{for } 1 \leq i \leq M \text{ and } 1 \leq i, j \leq N; \quad (1.1)$$

the edge e_s is then said to be *incident* on nodes v_i and v_j .

For a given set of nodes, $\mathcal{V}$, and a corresponding set of edges, $\mathcal{E}$, we write

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \quad (1.2)$$

to define the graph $\mathcal{G}$.

In general each edge may be associated with a *direction* and a *weight*, $w_i \in \mathbb{R}$. In a *directed graph* we attach a direction to each edge $e_s^{(d)}$. $e_s^{(d)} = (v_i, v_j)$ means that the edge e_i starts at node v_i and ends at node v_j . In an *undirected graph* the order in which nodes are written does not matter and $e_s^{(u)} = (v_i, v_j) = (v_j, v_i)$. Quite generally we allow for $v_i = v_j$, that is an edge may originate and end on the same vertex; this edge is said to form a *one-edged loop* attached to node v_i . It is also possible to allow more than one edge between nodes v_i and v_j .

If a graph contains neither multiple edges between pairs of nodes nor loops, then the graph is called *simple*. For simple graphs a number of additional statements can be made. For example, the number of edges in a simple graph is at most

$$M^{\max} = \frac{N(N-1)}{2}, \quad (1.3)$$

in which case the network is called *fully connected*.

Figure 1.1 shows an example of an undirected simple network with $N = 8$ nodes and $M = 7$ edges, and a directed network. Note that node 4 is disjoint from the rest of the network. While genes or proteins which do not interact with other molecules inside their environment are biologically implausible, it is nevertheless possible that,

for instance, a protein's interaction partners are not included in the experimental setup.

1.3.1.1. *Characteristics of a node*

Biological networks are generally *labelled* with information. To each node v_i we have an associated vector of properties, V_i . These may include the biological name of the node, e.g. the name of the gene or protein, biological classifications and other experimental data.

One of the most prominent characteristics of a node in a network is its *degree*, d_i , the number of edges incident on a node. In a directed network we distinguish between the *in-degree* and the *out-degree*, d_i^{in} and d_i^{out} , i.e. the number of nodes ending on and starting from node v_i .

The degree of a node tells us how many neighbours it has in the network. We define the *neighbourhood*, $\Gamma(v_i)$ of a node v_i through

$$\Gamma(v_i) := \{\nu_j | \nu_j \in V \text{ and } (\nu_i, \nu_j) \in E\}. \quad (1.4)$$

Trivially, the degree (in-degree) is also the size of the neighbourhood $d_i := |\Gamma(v_i)|$. In all networks we also have

$$\sum_i d_i = 2M \quad (1.5)$$

where $M = |E|$ is the total number of edges in a graph. (For directed networks the sum is M and not $2M$.) From Eqn. (1.5) it follows straightforwardly that the total number of nodes with odd degrees must be an even number.

1.3.1.2. *Paths, components and trees*

A *path* from node v_i to v_j is a sequence of edges which can be traversed to reach v_j starting from v_i ; in directed networks paths cannot go against the direction of an edge. We say that node v_j is *connected* to node v_i if there is a path from node v_i to v_j , taking into account the directionality of edges in a directed network. Thus node 1 in the network shown in Fig. 1.1B is connected to node 4; equally node 4 is connected to node 1. Node 2, however, is not connected to node 1. In an undirected network, if there is a path from node v_i to node v_j , then there is also a path from v_j to v_i . If there is a path starting from and ending on a node $v_i \in \mathcal{V}$, then this is called a *loop*.

A set of k nodes $\mathcal{C} = \{v_1, v_2, \dots, v_k\}$ where each node in $\mathcal{C}$ can be reached from other nodes in $\mathcal{C}$ but not from any node outside of $\mathcal{C}$ is called a *connected component of size k* of the network. In a simple network the number of components K is given by

$$K \geq N - M \quad (1.6)$$

which is easily shown by induction.

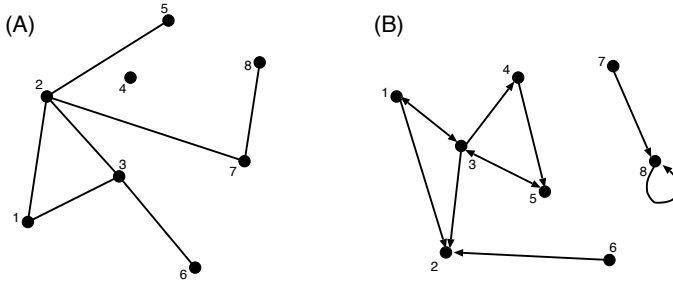


Fig. 1.1. Examples of a simple undirected network (A) and a directed network (B).

In many cases it may be preferable to study the largest connected component rather than the network as a whole. This may, for example, be the case when a large number of nodes occur in singletons, pairs or other small groups of nodes.

If there is more than one path between a pair of nodes $v_i, v_j \in \mathcal{V}$, then the graph contains *closed paths*, or loops. In an undirected simple graph, if there is precisely one path between each pair of nodes $v_i, v_j \in \mathcal{V}$, then there cannot be any loops and the graph is called a *tree*. If a graph consists of several components, each of which is a tree, the graph is sometimes referred to as a *forest*. The concept of a tree is very important and useful in the analysis of graphs and networks and we will sometimes borrow from the rich literature on trees.

Of particular interest is the *spanning tree* $\mathcal{T}$ of a connected graph with nodes $\mathcal{V}_{\mathcal{T}} = \mathcal{V}_{\mathcal{G}}$ and edges $\mathcal{E}_{\mathcal{T}} \subseteq \mathcal{E}_{\mathcal{G}}$, such that $(\mathcal{V}_{\mathcal{T}}, \mathcal{E}_{\mathcal{T}})$ is a tree. It is possible to show that a connected graph contains at least one spanning tree. Spanning trees can be used to traverse all nodes of a connected network.

1.3.1.3. Distance and diameter

If two nodes are connected by a sequence of nodes and edges, then the *distance* l_{ij} between them is defined as the number of edges that have to be traversed to reach node v_j from v_i ;

$$l_{ij} = \min\{X_{ij} | X_{ij} \text{ is a path from node } v_i \text{ to node } v_j \text{ along edges } e_s \in \mathcal{E}\}. \quad (1.7)$$

If there is no path by which node v_j can be reached from node v_i then we set

$$l_{ij} = \infty. \quad (1.8)$$

In directed networks, of course l_{ij} can be different from l_{ji} ; one of them can even be infinite as shown by nodes 1 and 2 in the network in Fig. 1.1 where $l_{12} = 1$ and $l_{21} = \infty$.

The *diameter* of a network is defined as the maximum distance between two nodes in the network,

$$D = \max\{l_{ij} | v_i, v_j \in \mathcal{V}\}. \quad (1.9)$$

Thus by definition the diameter of the network which consists of more than one component is ∞ . The definition for D is analogous to the definition of diameters in geometry and topology: the maximum distance between two points belonging to the same object.

Frequently, we therefore restrict analyses of biological networks to the nodes in the largest component. This is particularly relevant if the network exhibits a *giant connected component* (GCC) which is defined for growing networks only. A GCC is a component with non-zero relative size as the size of the network becomes large. The relative size of a component is defined as the number of nodes in the component divided by the total number of non-zero degree nodes. Because of the incomplete nature of many biological data sets, observed biological networks often appear fragmented and composed of several components. However, once a complete or truly integrated network, one which contains all physical, regulatory and small-molecule-mediated interactions has been established, we would expect all the nodes in the whole network to be connected.

1.3.2. Network properties

Some of the quantities introduced above can be used to characterise aspects of networks. Here we will introduce some of the common statistics that have been used to describe them.

1.3.2.1. The degree distribution

We have already discussed the degree of a node v_i , here denoted by d_i . The average degree, $\bar{d}$, of a network is given by

$$\bar{d} = \frac{1}{N} \sum_{i=1}^N d_i. \quad (1.10)$$

We note that in a directed network the average in- and out-degrees of a node must be equal,

$$\frac{1}{N} \sum_{i=1}^N d_i^{\text{in}} = \frac{1}{N} \sum_{i=1}^N d_i^{\text{out}}. \quad (1.11)$$

Surprisingly, this simple fact is frequently ignored and any analysis which contains reports of unequal in- and out-degrees should be treated with considerable caution.

The degree is analogous to the coordination number of a site in a regular lattice. Unlike coordination numbers, however, the degrees of nodes in a network will generally take on many different values. Thus the average degree is not very informative about a network and what is generally considered instead, is the degree distribution $n(k)$, the probability of a node to have degree $d_i = k$, $k = 0, 1, 2, \dots$.

The degree distribution is defined by

$$n(k) = \frac{1}{N} \sum_{i=1}^N \delta_{d_i, k} \quad \text{for } k = 0, 1, 2, \dots \quad (1.12)$$

where $\delta_{i,j}$ is the Kronecker delta function

$$\delta_{i,j} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \quad (1.13)$$

defined for integers i, j . The degree distribution summarises information about the local environments in a network. It has to be kept in mind, though, that the degree distribution is highly degenerate, i.e. there are many different networks which have the same degree distribution. While the average in- and out-degrees in networks have to be identical, the corresponding degree distributions,

$$n^{\text{in}}(k) = \frac{1}{N} \sum_{i=1}^N \delta_{d_i^{\text{in}}, k} \quad (1.14)$$

and

$$n^{\text{out}}(k) = \frac{1}{N} \sum_{i=1}^N \delta_{d_i^{\text{out}}, k}, \quad (1.15)$$

respectively, can be very different indeed.

1.3.2.2. Clustering

A further statistic which describes the local environment, but also including next-nearest neighbours, is given by the so-called clustering coefficient. The clustering coefficient measures the probability that two nodes v_j and v_k , which are both neighbours of v_i (i.e. $(v_i, v_j), (v_i, v_k) \in \mathcal{E}$ in an undirected graph), are themselves connected by an edge $(v_j, v_k) \in \mathcal{E}$. For node v_i the clustering coefficient is defined by

$$c_i = \frac{2\eta_i}{d_i(d_i - 1)} \quad \text{for } d_i \geq 2 \quad (1.16)$$

where η_i is the number of edges among the nodes connected to v_i . The average clustering coefficient of the network is then given by

$$\bar{c} = \frac{1}{N} \sum_{i=1}^N c_i. \quad (1.17)$$

In a social network the clustering coefficient could for instance measure the extent to which my friends are also friends themselves.

Just like the average degree fails to capture the diversity of degrees observed in most natural networks, the average clustering coefficient fails to describe the

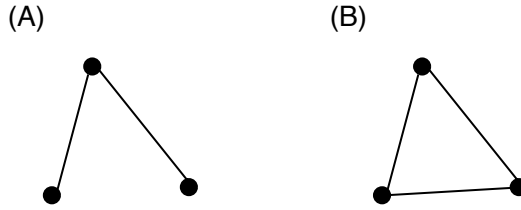


Fig. 1.2. Three connected nodes in an undirected network can either form an open (A) or a closed triangle (B). A network's transitivity is defined as the probability of a triangle to be closed on all three sides.

network's local inhomogeneity. It is therefore often useful to study the distribution of clustering coefficients, e.g. using the cumulative distribution defined by

$$C(c) = \sum_{i=1}^N \int_0^c \delta(c_i - c') dc' \quad (1.18)$$

where $\delta(x)$ is the Dirac delta function, defined by $\delta(x) = 1$ for $x = 0$ and $\delta(x) = 0$ otherwise.

Related but not identical to the clustering coefficient is the transitivity. This is defined by

$$T = \frac{\# \text{ of closed triangles}}{\# \text{ of connected triplets of nodes}}. \quad (1.19)$$

For trees we necessarily have $\bar{c} = 0$; the same is also true for the square (or cubic or hypercubic lattices). Thus small values of C are not indicative of the absence of loops or closed paths. In fact, as we shall see later, most naturally occurring lattices, including those in systems biology, are locally tree-like. For this reason we prefer the distribution of clustering coefficients rather than the average clustering coefficient.

1.3.2.3. Average path length

The average path length of a network follows from all pairwise distances in a network and is given by

$$\bar{l} = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N l_{ij}. \quad (1.20)$$

By definition $l_{ii} = 0$.

Analogous to the degree and clustering distributions, it is also possible to define a distribution of network distances. One convenient definition is given by

$$\lambda(l) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=1}^N \delta_{l_{ij}, l} \quad \text{for } l = 1, 2, \dots, \quad (1.21)$$

which counts the number of distances of length l .

Because the distance of two unconnected nodes is ∞ , the average path length (and the diameter) will diverge in networks which consist of more than one component. Therefore one often considers only the largest connected component when analysing network distances. We note that the diameter D and the average path length in a network may be very different.

1.3.3. Mathematical representation of networks

There are three basic methods to represent or store a graph. Here we will define these different representations before giving some guidelines on when to use which representation.

1.3.3.1. The adjacency matrix

The *adjacency matrix* $\mathbb{A}$ of a graph is an $N \times N$ matrix and is defined by

$$A_{ij} = \begin{cases} w_{ij}, & \text{if nodes } i \text{ and } j \text{ are connected by an edge with weight } w_{ij} \\ 0, & \text{otherwise.} \end{cases} \quad (1.22)$$

This is the most general case but we will often consider special cases of Eqn. (1.22). For an unweighted graph, for example, $w_{ij} = n_{ij} \in \mathbb{Z}_0$ is the number of (directed) edges between nodes v_i and v_j . For an undirected graph we have

$$A_{ij} = A_{ji}, \quad (1.23)$$

i.e. the adjacency matrix is symmetrical. The adjacency matrix of a *simple* graph is given by

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between node } i \text{ and } j \text{ and } j \neq i \\ 0 & \text{otherwise.} \end{cases} \quad (1.24)$$

For real networks, as we will see below, the actual number of edges is much lower than the maximum number of edges possible, Eqn. (1.3), and the adjacency matrix will be a *sparse matrix*.

The adjacency matrix of the simple undirected graph in Fig. 1.1, for example, is given by

$$\mathbb{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad (1.25)$$

Table 1.1. Computational complexity of some elementary graph operations in terms of the number of nodes, N , and number of edges, M . Costs also include a constant factor which has been ignored here.

Property	Adjacency matrix	Adjacency list	Edge list
Memory requirement	N^2	$N + M$	M
Initialisation	N^2	N	1
Copying a node	N^2	M	M
Deleting an edge	N	M	1
Finding an edge	1	N	M
Is a node isolated	N	1	M
Testing for a path between two nodes	N^2	$M \log(N)$	$N + M$

where the nodes and columns correspond to the node labels in Fig. 1.1. The labelling of the nodes can of course be changed and the corresponding new adjacency matrix can be obtained from the adjacency matrix in Eqn. (1.25) by rearranging the rows and columns.

1.3.3.2. The adjacency list

We see in Eqn. (1.25) that the adjacency matrix is sparse. This is typical for many real networks and the adjacency matrix will typically have only a small fraction of non-zero entries. An alternative and slightly less wasteful way of storing the structure of the network is through the *adjacency list*. This list contains all nodes connected to a node; the adjacency list corresponding to the matrix in Eqn. (1.25) is

$$\begin{aligned}
 &1 : 2, 3 \\
 &2 : 1, 3, 5 \\
 &3 : 1, 2, 6 \\
 &4 : \\
 &5 : 2 \\
 &7 : 2, 8 \\
 &8 : 7
 \end{aligned} \tag{1.26}$$

Computationally this is generally implemented by defining an array of lists such that the nodes connected to a given node can be accessed immediately.

1.3.3.3. The edge list

The two representations introduced above focus on nodes. In some instances it may be more interesting to describe the edges, e.g. when we want to study if two

interacting biological molecules share the certain characteristics. In this case we can use the *edge list* notation. This, for the above example, takes the form

$$\{(1, 2), (1, 3), (2, 3), (2, 5), (2, 7), (3, 6), (7, 8)\}. \quad (1.27)$$

Thus we store a list containing each edge that exists in the graph, keeping in mind that for an undirected graph $(v_i, v_j) = (v_j, v_i)$. In many circumstances the edge list is the most memory-efficient way to store network information.

1.3.3.4. *Some remarks on complexity*

Here, *complexity* refers to the computational effort required to evaluate a property of the graph. The effort of performing simple computational tasks such as setting up a network or testing if two nodes are connected depends on the way in which network information is represented. The complexities of a number of different tasks for the three network representations outlined above are given in Table 1.1. Strictly speaking, the true cost of each task is proportional to the factor in Table 1.1 multiplied by a constant factor.

All real networks are finite sized and, as far as biological networks are concerned, *mesoscopic systems*. The number of nodes is typically of the order of several thousand to tens of thousands. This implies that (i) in principle, it is possible to analyse networks computationally and (ii) the size of the network is sometimes of the same order as the proportionality constant by which the complexities in Table 1.1 are multiplied.

The computational complexity of several important and interesting problems in the analysis of networks belong, however, to classes of problems which are considerably more cumbersome. Briefly, problems are often divided into the following classes

P: A problem that can be solved in polynomial time.

NP: (Non-deterministic polynomial) A problem that has a solution that can be verified (by a non-deterministic Turing machine) in polynomial time. All problems in *P* are also in *NP*; the reverse is not necessarily true.

NP-hard: A problem that can be solved by an algorithm which can be translated into one for solving any other *NP* problem. *NP-hard* problems are at least as hard to solve as any other problem in *NP*.

NP-complete: A problems that is both in *NP* and *NP-hard*.

Issues of computational complexity are frequently encountered in the analysis of networks. Especially when trying to understand properties of theoretical network models or when assessing statistical significance of network properties, we will often have to repeatedly calculate the same network property.

1.4. Comparing Biological Networks

In the previous section we have discussed some basic mathematical properties of networks. Unfortunately, as will be discussed later, networks with identical/similar properties are not necessarily identical/similar. Moreover it has so far been impossible to come up with a useful definition of *distance* between networks. Here, we therefore only briefly discuss basic notions of network identity as far as these are required in order to compare biological networks.

Comparative analysis is a cornerstone of evolutionary analysis and at the sequence level has provided us with detailed insights into the evolutionary history of life. Thus the biological analysis of networks must necessarily involve comparison of networks from different species. For example there has been considerable interest as to whether evolutionary inferences from protein interaction network data provide similar information in different organisms. But while the vagaries of the highly stochastic evolutionary process are already hard enough to understand at the level of DNA and protein sequences, these problems are exacerbated at a spectacular scale once we enter the system level. Here we therefore focus only on the basics of the underlying theoretical framework that may aid in comparing biological networks.

An important lesson that can be learned from sequence-based (or even traditional morphological-trait-based) comparative biology is the need to compare species over the broadest range of evolutionary divergences possible. Our understanding of sequence evolution (including the evolution of e.g. transcription factor binding sites) has benefited enormously from the abundance of data from several closely related species. For many biological networks, the evolutionary separation between model organisms is simply too large for meaningful comparisons to be made. We therefore need to map interactomes, gene regulatory and metabolic networks in those species that are sufficiently closely related to model species such as *S. cerevisiae* and *E. coli*.

1.4.1. Identity of networks

Two networks $\mathcal{G}_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2 = (\mathcal{V}_2, \mathcal{E}_2)$ are called *isomorphic* if there is a one-to-one correspondence between the nodes, $\mathcal{V}_1$ and $\mathcal{V}_2$, and edges, $\mathcal{E}_1$ and $\mathcal{E}_2$, which preserves the assignment of nodes to edges and *vice versa*. That is, if $e_s \in \mathcal{E}_1$ is associated with $e_t \in \mathcal{E}_2$, and if $e_s = (v_i, v_j)$ and $e_t = (v_k, v_l)$, then v_i must be associated with v_k and v_j with v_l .

If $\mathcal{G}_1$ and $\mathcal{G}_2$ are isomorphic we write

$$\mathcal{G}_1 \simeq \mathcal{G}_2 \tag{1.28}$$

rather than $\mathcal{G}_1 = \mathcal{G}_2$ to indicate that $\mathcal{G}_1$ and $\mathcal{G}_2$ are instances of the same (abstract) graph; they may still have different graphical or mathematical representations: for

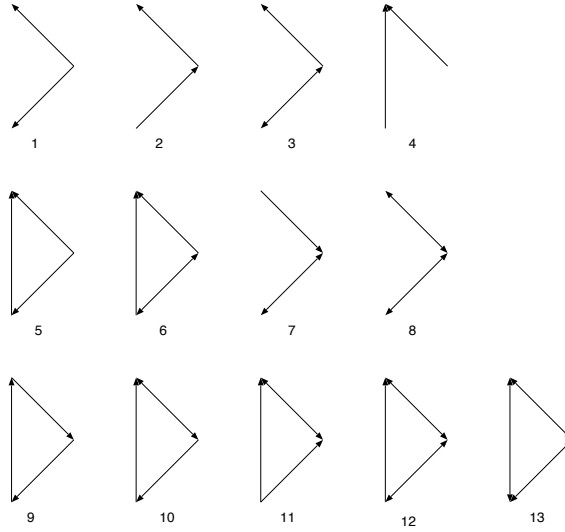


Fig. 1.3. The 13 patterns possible to observe for three connected nodes in a directed networks.

example, the rows or columns of their respective adjacency matrices may be interchanged.

Each network can be drawn in many different ways. We also say that a graphical representation of a network is an *instance* of a network and we will seek to define under what circumstances two networks are identical, in the sense that their network structure is the same.

Determining if two graphs are isomorphic has been shown not to be in P but so far there has been no proof that it is *NP-complete*. Some people prefer to assign it to its own class of *graph isomorphism problems*. In practice, these issues may pose severe limitations on the exhaustive analysis of biological networks. For example, a human protein-interaction network which covers the 20,000 or so different proteins (ignoring splice variants) cannot easily be analysed in a comprehensive statistical manner. For computational reasons the search for suitable heuristics for network investigation will therefore increase in importance.

1.4.2. Subnets and patterns

A *subnet* $\mathcal{S}$ of a network $\mathcal{N}$ is defined by $\mathcal{S} := (\mathcal{V}^*, \mathcal{E}^*)$ with

$$\mathcal{V}^* \subset \mathcal{V}$$

$$\mathcal{E}^* \subset \mathcal{E}$$

$$\text{If } e_s = (v_i, v_j) \in \mathcal{E}^* \text{ then } v_i, v_j \in \mathcal{V}^*$$

$$\text{If } v_i, v_j \in \mathcal{V}^* \text{ and } (v_i, v_j) \in \mathcal{E} \text{ then } e_s = (v_i, v_j) \in \mathcal{E}^* \quad (1.29)$$

Thus a subnet is itself a network consisting of a subset of nodes of the *global network* $\mathcal{G}$ and all the edges connecting pairs of nodes in the subnet. Equally, we could define the subnet through the set of edges and the associated nodes.

The way subgraphs are set up can influence the inferences to be gained from an analysis of $\mathcal{S}$. We may, for example, study a particular biochemical pathway as a subset of an organism's metabolism; or we may seek to test for interactions among the known proteins in an organism.

Closely related to subnets is the notion of a *pattern* which we define through a connected graph $\mathcal{P} := (\mathcal{V}_{\mathcal{P}}, \mathcal{E}_{\mathcal{P}})$; we define the size of the pattern as the number of nodes needed to define it, $s = |\mathcal{V}_{\mathcal{P}}|$. For example, nodes 1, 2 and 3 in Fig. 1.1A form a closed triangle which is a pattern of size 3. In many cases we will be interested in determining the frequencies of a set of patterns in a network. The sets of all patterns formed by three nodes in a directed network are shown in Fig. 1.3; the corresponding patterns of size 3 in an undirected network are in Fig. 1.2. These patterns may represent important functional or logical units of organisation; of particular interest are those patterns in a network which have more internal edges than would be expected to occur by chance, given the rest of the network.

1.4.3. *The challenges of the data*

We have already mentioned the complexity of evolutionary processes, especially when trying to go beyond the sequence level. The analysis of this highly stochastic and contingent process is exacerbated when one considers the often woeful quality of the data: for protein interaction networks (PIN) the rates for false-positive and false-negative results are estimated to be around 40%. Bioinformatics and statistics may help to clean the data to some extent but improvements in experimental techniques offer the only real solution to this problem. Although important and interesting we will here not be concerned with such issues of quality control. Rather we will discuss what should be included in theoretical descriptions of complex networks in a biological setting.

It has to be kept in mind, though, that present network data are highly averaged and artificial constructs: the language of graph theory may simply be too static to usefully describe complex biological networks. We may in approximation seek to understand networks as entities that change over three different time scales: (i) they will change over evolutionary time scales between species (millions of years), (ii) they will change during the course of an organism's development (years), and finally, (iii) connections will be formed and lost in response to physiological change and external stimuli (sub-second to minutes). Already we are seeing the first attempts to map biological networks *in vivo* and future experimental developments will, no doubt, enable us to probe the dynamics on the biologically relevant time and spatial scale. For protein interaction networks, experimental methods can at the moment only resolve the changes in PIN structure accumulated between species,^{16–18} but the data are not yet sufficiently reliable to make meaningful comparisons.

References

1. P. Uetz, L. Giot, L. G. Cagney, T. Mansfield, R. Judson, V.D.L. Narayan, M. Srinivasan, P. Pochart, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields and J. Rothberg A comprehensive analysis of protein-protein interaction networks in *saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.
2. S. Maslov and K. Sneppen Specificity and stability in topology of protein networks. *Science*, 296(5569):910–3, 2002.
3. I. Agraftioti, J. Swire, J. Abbott, D. Huntley, S. Butcher and M.P.H. Stumpf Comparative analysis of the *saccharomyces cerevisiae* and *caenorhabditis elegans* protein interaction networks. *BMC Evolutionary Biology*, 5:23, 2005.
4. H. Ma and A.P. Zeng Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19:270–277, 2003.
5. M. Ronen, R. Rosenberg, B. Shraiman and U. Alon Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA*, 99(16):10555–10560, 2002.
6. A. Evangelisti and A. Wagner Molecular evolution in the yeast transcriptional regulation network. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, 302B(4):392–411, 2004.
7. R. Albert and A.L. Barabasi Statistical mechanics of complex networks. *Rev.Mod.Phys.*, 74(1):47–97, 2002.
8. M. Newman The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
9. T. Evans Complex networks. *Contemporary Physics*, 45(6):455–474, 2004.
10. S. Dorogovtsev and J. Mendes *Evolution of Networks*. Oxford University Press, 2003.
11. M.P.H. Stumpf, W.P. Kelly, T. Thorne and C. Wiuf Evolution at the system level: the natural history of protein interaction networks. *Trends Ecol.Evol.*, 22:366–373, 2007.
12. A.P. Coates, S.H. Muggleton and M.J.E. Sternberg The identification of similarities between biological networks: Application to the metabolome and interactome. *Journal of Molecular Biology*, 369:1126–1139, 2007.
13. S. Proulx, D. Promislov and P. Phillips Network thinking in ecology and evolution. *Trends.Ecol.Evol.*, 20(6):345–353, 2005.
14. R.M. May Network structure and the biology of populations. *Trends.Ecol.Evol.*, 21:394–399, 2006.
15. G. Robins and P. Pattison Random graph models for temporal processes in social networks. *J.Math.Soc.*, 25:4–21, 2001.
16. H.B. Fraser, A.E. Hirsh, L.M. Steinmetz, C. Scharfe and M.W. Feldman Evolutionary rate in the protein interaction network. *Science*, 296(5568):750–2, 2002.
17. I.K. Jordan, Y.I. Wolf and E.V. Koonin No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol Biol*, 3(1):1, 2003.
18. H. Qin, H.H.S. Lu, W.B. Wu and W.H. Li Evolution of the yeast protein interaction network. *Proc. Natl. Acad. Sci. USA*, 100(22):12820–4, 2003.

This page intentionally left blank

Chapter 2

Evolutionary Analysis of Protein Interaction Networks

CARSTEN WIUF¹ AND OLIVER RATMANN²

¹*Bioinformatics Research Center, Aarhus University*

²*Centre for Biostatistics, Imperial College London*

wiuf@birc.au.dk, oliver.ratmann@imperial.ac.uk

Systems approaches to understanding the structure, organisation and functioning of organisms and cells are now becoming commonplace. In this chapter we focus on protein interaction networks and their potential use for inference on the evolutionary processes that have shaped the interactome, the collection of all proteins in a cell together with their physical interactions. We demonstrate that simple mathematical models may capture essential aspects of the processes and use these to develop a Bayesian likelihood-free scheme for inference on three small organisms *T. pallidum*, *H. pylori* and *P. falciparum*.

2.1. Introduction

Postgenomic data such as protein interaction networks (PINs) or regulatory networks offer a new reflection on the interactome, here defined as the entire collection of all proteins in a cell or organism together with their interactions, and may be used in addition to individual gene or genomic approaches to elucidate the evolution of living systems across the tree of life.^{1,2} PINs are incomplete observations of the interactome and can be described as a graph which contains a set of nodes, interacting proteins and edges, the observed interactions between the proteins, whereas regulatory networks consist largely of the functional linkages among regulatory genes that produce transcription factors, and their target cis-regulatory systems of other regulatory genes. On the network level, extensive variation and evolutionary conservation has been identified,^{3–6} leading our understanding of the evolution of biological networks into uncharted terrain.^{7,8} In the context of protein network evolution, a number of processes motivated from molecular genetic data are being studied^{9–14} and gene duplication is sought to have a key role in network evolution across domains,¹⁵ perhaps with an even greater role in eukaryotes than prokaryotes.¹⁶

This chapter aims at describing some recent advances in mathematical modeling and statistical analysis of network data, with emphasis and applications to an *evolutionary* analysis of PIN datasets. Data should be analysed using models that

adequately describe the data *and* the mechanisms generating it. Models should be as simple as possible, but not simplistic in that realistic extensions to the model alter the data analysis fundamentally. We will develop models of network growth that may qualitatively explain the topology of observed PIN datasets and mimic key forces in biological evolution. We will demonstrate how likelihood-free inference (LFI) affords to statistically analyse these models of network growth in extensive computer simulations. Caution is warranted in the interpretation of the results without a full understanding of these models, and we will investigate simple, topological patterns under these models with full mathematical rigour. Taken together, these provide insight into the broad dynamics of network evolution.

A myriad of physical mechanisms may contribute to the evolution of the interactome, and their relative roles in network evolution for different species in different population genetic environments remain unclear. We begin with a brief overview.

2.1.1. *Molecular genetic uptake*

The phylogenetic relation of the major bacterial lineages does not seem to emerge reliably, suggesting rapid evolution of each lineage and/or formidable rates of lateral gene transfer.¹⁷ The genomic mechanisms of lateral gene transfer include molecular genetic uptake through conjugation, transduction, transformation, gene transfer agents and gene loss.¹⁸ The mechanisms by which networks evolve under such molecular uptake remain unclear but see Fig. 2.1 for possible modes of evolution. A recent study of *E. coli* suggests that its metabolome evolves by direct uptake of peripheral reactions in response to changed environments.¹⁹ Recent comprehensive analyses across 181 prokaryotic genomes suggest that lateral gene transfer probably occurs at a low rate, but that cumulatively, about 80% of all genes in a prokaryotic genome are involved in lateral gene transfer, and once acquired, are then vertically transferred.²⁰

2.1.2. *Expansion by gene duplication*

The importance of gene duplication to biological evolution has long been recognised and substantial evidence elucidating the importance and the mechanisms of this process in higher organisms has been collected from genomic sequence data.^{21,22} Genes duplicate at rates of 0.1–1% per generation per haploid genome.²³ The molecular mechanisms by which duplicate genes arise are diverse, ranging from whole genome duplication (WGD) to more restricted duplications of chromosomal regions.²³ Of the latter, single gene duplications (SGD; see Fig. 2.1) appear to occur most often; in *C. elegans*, for example, only $\approx 50\%$ of duplicated regions appear to be long enough to contain a complete gene on average. Just after a successful SGD, the child and the parental gene products have exactly the same functions and protein interactions, but over a relatively short evolutionary time,²³ the two genes may assume one of several fates: (D1) one gene may be silenced

(non-functionalisation), (D2) both genes are preserved such that one is functionally redundant to the other, (D3) both genes acquire mutually exclusive deleterious mutations (sub-functionalisation), or (D4) one gene may acquire a new function while the function of the other is retained (neo-functionalisation).

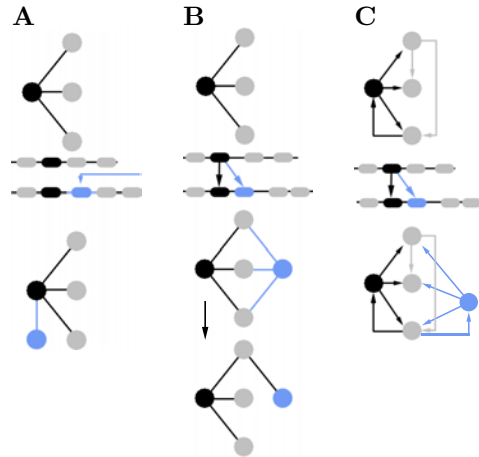


Fig. 2.1. Top-down schema representing possible modes of protein and regulatory network evolution. (A) Protein interaction network before and after lateral gene transfer (blue). (B) Protein interaction network before and after a successful, single tandem gene duplication, with the new, fixed duplicate depicted in blue. (C) Regulatory network before and after a successful tandem duplication of a transcription factor.

D3 does not rely on the sparse occurrence of beneficial mutations, but on loss-of-function mutations in regulatory regions; this is very attractive because it might explain the abundance of retained duplicates and the emergence of molecular genetic incompatibilities in allopatric subpopulations of a species. Indirect evidence also suggests that D3 may frequently occur not only in multicellular organisms, but also in unicellular species such as those under study.²³ Importantly, various lines of evidence suggest that protein interactions derived from gene duplicates may persist over evolutionary time scales.^{24,25} In the three species we use here, *H. pylori*, *T. pallidum* and *P. falciparum*, there is no recorded evidence of WGD and we will simply focus on SGDs in the following discussion, though we note that for other species such as *S. cerevisiae*, WGDs have played an important role.²³

2.1.3. Redeployment of existing genetic systems

More recently, the alteration of genetic regulatory systems has come under intensive study.^{4,26,27} Considering closely related species, remarkable evolutionary plasticity and conservation has been identified for a number of subnetworks,²⁷ providing a first insight into the mechanisms underlying the evolution of regulatory networks. While these networks may evolve by gene duplication,²⁸ we here point out the quali-

tative difference that relatively small regulatory changes may result in extraordinary modifications of the interactome, such as the redeployment of entire genetic systems displayed in Fig. 2.1.²⁷

2.2. Protein Interaction Network Data

A number of PIN datasets are now available for both the prokaryotic and eukaryotic domains.^{29–38} These have been compiled by a variety of high-throughput techniques, most prominently yeast two-hybrid systems and tandem affinity purification,³⁹ and may be augmented with literature-curated and/or computationally inferred interactions. These datasets provide at least a static picture of protein interactions that may occur under one or a defined set of *in vivo* conditions.

PIN datasets are flawed with a number of shortcomings, most prominently high levels of noise⁴⁰ and incompleteness.⁴¹ In reality, the subset of interactions that has been experimentally identified is not random, either because not all proteins are known, or the experimenter might choose to work with a subset of the known proteins only, or the experimental technique is not suitable to identify all existing interactions equally well. Interactions are often validated by multiple occurrence across independent experiments; this increases the reliability of individual interactions, but may add further sampling bias to the dataset.⁴² Here, we consider binary, undirected high-confidence interactions derived from multiple validation; Table 2.1 lists some examples, including the three organisms we are analysing, the eukaryote *P. falciparum*, and the bacteria *T. pallidum* and *H. pylori*. The question, whether current PIN datasets are representative of the transient, temporarily and spatially heterogeneous interactome is further fuelled by the fact that datasets are highly averaged: not only over technical aspects such as the experimental protocol, but also over interaction strength, between individual variation and the precise cellular conditions under which interactions take place. The latter is particularly problematic for multicellular organisms; here we focus on the network evolution of some unicellular organisms.

Nevertheless, PIN datasets are increasingly useful for elucidating the evolution of living systems;^{12,14,43,44} we ask here if and how the topology of PIN datasets may help to understand the evolution of the interactome of unicellular organisms. We take a practical approach, regarding PIN datasets as single, co-dependent observations, which are at present and as a whole devoid of important population characteristics,⁸ and pay particular attention to missing data.

2.3. Mathematical Models of Networks and Network Growth

With the first available experimental PIN datasets, it became apparent that real networks have some very different properties from the canonical mathematical descriptions of networks, such as random graphs or regular lattices.⁴⁵ This sparked consid-

Table 2.1. PIN Datasets.^a

	Organism	Proteins ^b	Interactions ^c	Genes ^d	In % ^e
Prokaryotes	<i>T. pallidum</i> ²⁹	575	978	1,039	55
	<i>H. pylori</i> ³⁰	675	1,096	1,500	45
	<i>C. jejuni</i> ³¹	1,047	2,668	1,884	56
	<i>M. loti</i> ³²	1,607	2,079	6,750	24
	<i>E. coli</i> ³³	1,852	6,976	4,290	43
	<i>C. synechocystis</i> ³⁴	1,917	3,211	4,003	48
Eukaryotes	<i>P. falciparum</i> ³⁵	1,271	2,642	5,300	24
	<i>C. elegans</i> ³⁶	2,638	3,970	22,000	12
	<i>S. cerevisiae</i> ³⁷	4,013	10,056	5,500	73
	<i>D. melanogaster</i> ³⁸	7,451	22,636	12,900	58

^aAvailable PIN datasets, in relation to the unknown interactome. Protein interaction databases such as IntAct (<http://www.ebi.ac.uk/intact/>) provide information on available PIN datasets.

^bNumber of proteins for which reliable interaction data was obtained.

^cNumber of experimentally observed interactions; for details of the high-confidence sets, we refer to the literature as indicated. Self-interactions are removed.

^dEstimated number of open reading frames (ORFs) in the respective genome.

^eSampling fraction, Nodes/Genes.

erable interest in describing aspects of networks, such as the degree sequence,⁴⁶ and classifying networks according to some of its features, most notably the profile of subnetwork (motif) occurrences.⁴⁷ More recently, interest has shifted towards models of network growth, with PIN datasets assuming a secondary role that, among others, may inform the evolutionary history of the interactome. The complexity of the problem however comes at a price: analysing models of interactome evolution is intimately linked to a development of novel computational methods.

2.3.1. *Simplistic models of network growth*

Many of the descriptive approaches to understanding aspects of cellular organisation are implicitly based on network models that are evolutionarily implausible. To analyse the significance of features of network data, null datasets are commonly generated from the observed network by randomising the nodes and keeping particular aspects of the network fixed. The most popular rewiring procedure keeps the node degree distribution fixed and redistributes the links between proteins. This rewiring procedure is tempting as a null model for testing hypotheses about the observed data, since it is easy to use and falsely suggests goodness of fit by keeping a *single* aspect of real networks fixed. Analogous parametric models exist, such as Exponential Random Graph Models (ERGM),^{48,49} a special case of which is the Erdős–Rényi (ER) graph.⁴⁵ An ER graph has a fixed number of nodes N , and each pair of non-identical nodes is connected with probability p . If N is large and p small, then the degree sequence is approximately Poisson with intensity $\lambda = Np$. Like all ERGM graphs, the above rewiring model generates networks where mo-

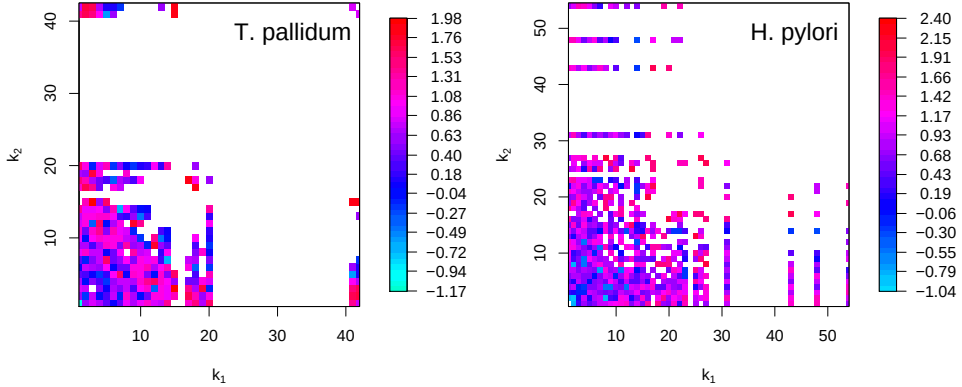


Fig. 2.2. Relative log connectivity distribution CONN (see Table 2.2) of the *T. pallidum*, *H. pylori* and *M. loti* PIN datasets. Deviations from zero (blue is zero) indicate departures from the homogeneous network with the same node degree distribution.

tifs are expected to be equally spread (homogeneous) throughout the network, in contrast to real PIN datasets; see Fig. 2.2. Taken together, the above rewiring procedure implicitly assumes a model of network growth that falls short in explaining key topological aspects of PIN datasets. In addition, such models have limited value in that neither p nor λ have an evolutionary interpretation and the biological importance of one value of p , or λ , rather than another might be difficult to assess.

Considering the descriptive analysis of network data, some progress is possible when several carefully chosen aspects of the observed network are kept fixed. However, a certain arbitrariness in choosing invariant aspects of the network cannot be avoided, and conditioning on different invariant aspects of PINs typically leads to different biological conclusions.⁵⁰

2.3.2. Complex models of network growth by repeated node addition

A number of mechanistic models have been proposed in biology and elsewhere to model network growth from a topological perspective. What these models have in common is to generate a network by gradually adding nodes and modifying, adding, or deleting links to a small initial graph. Collectively, these models are referred to as Randomly Grown Graphs (RGGs).^{43,51}

In a seminal paper, Barabási and Albert⁴⁶ found that many different naturally occurring networks exhibit a power-law degree distribution, and that a simple growth mechanism that locally modifies the network structure may roughly explain the shape of the degree distribution. Their model proceeds by repeating:

PA Choose m nodes with probability proportional to their degrees and introduce a new node. Add m links between the chosen nodes and the new node;

see Ref. 52 for a rigorous mathematical treatment. However, once m is fixed, PA is unable to generate certain classes of topological patterns; for example, PA with $m = 1$ generates only tree-like networks. Inspired by the important insight that network features may be explicable by simple rules, other RGGs that mimic evolutionary processes more closely *and* are able to create complex topological patterns that occur in real networks have been formulated.⁴³

Formally, RGGs are instances of Markov chains in the sense that the graph $\mathcal{G}_{t+1} = (\mathcal{V}_{t+1}, \mathcal{E}_{t+1})$ at step $t + 1$ only depends on the graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ at step t . We have already seen two (albeit unrealistic) examples, PA and the ER graph:

ER Introduce a new node and connect the new node to the existing nodes, each with probability p .

The structure of PINs derives from multiple stochastic processes over evolutionary time scales, so that it appears plausible to combine a number of growth mechanisms to model protein network topologies more realistically. The design of these mixture models depends on the biological problem in view. We ask here if the network topology provides any clues on whether gene duplication is likely to play a larger role in network evolution of eukaryotes than prokaryotes. One straightforward approach is to devise a two-component model, where one component models duplication and divergence (DD), and the other captures aspects of network growth which are not specifically related to D1–D3. Model PA has been applied to a variety of networks from theoretical physics, technology, and sociology; we here take it as a proxy for generic network growth. Assume a graph at step t , then at step $t + 1$ do PA as above with probability α and $m = 1$, or with probability $1 - \alpha$,

DD Choose a node v_{old} at random in $\mathcal{G}_t$ and introduce a new node v_{new} . For each neighbour v of v_{old} , create a link between v_{new} and v with probability p ; otherwise with probability r erase the link (v_{old}, v) and create the link (v_{new}, v) . Create a link between v_{old} and v_{new} with probability q .^a

Model DD+PA is illustrated in Fig. 2.3. Here, we fix $r = 0.5$, i.e. the links (v_{old}, v) and (v_{new}, v) are equally likely; it has been argued that $r \neq 0.5$,⁹ but to date biological evidence for $r \neq 0.5$ appears to be inconclusive, see Ref. 23, p.225. More importantly, corresponding to the preservation of ancestral function(s), all links of v_{old} are maintained in the sense that at least one of the links (v_{old}, v) and (v_{new}, v) is present in $\mathcal{G}_{t+1}$ whenever v is a neighbour of v_{old} in $\mathcal{G}_t$.

The probability of a node of degree k under PA reaches $\text{Prob}(D = k|\text{PA}) = 4/(k(k+1)(k+2))$ in a large network, which asymptotically is a power-law.⁵¹ For

^aSee the discussion after Theorem 2.4 for technical modifications, which we apply in analysis of data.

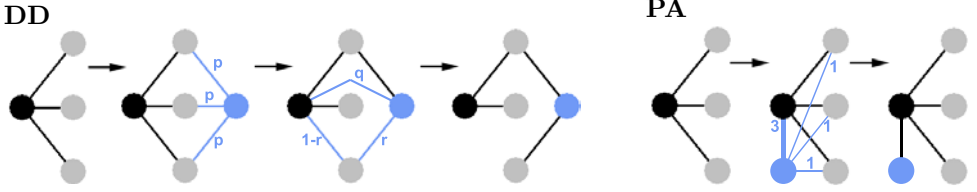


Fig. 2.3. Schema of network growth by model DD+PA; at each step of node addition, mechanism PA is chosen with probability α , and mechanism DD is chosen with probability $1 - \alpha$ as detailed in the main text.

the mixture model, our intuition may be fostered in a similar vein, as detailed in the next section.

2.3.3. Asymptotics of the node degree DD+RA and DD+PA

Asymptotic statements about the degree distribution can be obtained for some mixture models, including DD+PA; we present here a subset of these results.^{53,54} These provide some qualitative insight into the properties of networks evolving under such models, aiding in their interpretation.

For a more stringent mathematical analysis, we will first replace the PA component with random attachment (RA);⁵⁴ with probability α ,

RA Choose a node v_{old} at random in $\mathcal{G}_t$ and introduce a new node v_{new} . Create a link between v_{old} and v_{new} .

The difference between the two growth mechanisms DA and RA is clear in terms of the node degrees. In contrast to PA, the degree distribution is geometric $\text{Prob}(D = k | \text{RA}) = 2^{-k}$ under model RA.⁵³ Under DD+RA, the expected number, $n_t(k)$, of nodes with degree k fulfils the following recursion – called the *master equation* – for $t \geq t_0$, where t_0 is the size of the initial network:

$$\begin{aligned}
 n_{t+1}(k) = (1 - \alpha) & \left\{ \left(1 - \frac{1 + kp}{t} \right) n_t(k) + \frac{(k-1)p}{t} n_t(k-1) + (1-q)F_t(1-\phi, k) \right. \\
 & \left. + qF_t(1-\phi, k-1) + (1-q)F_t(p+\phi, k) + qF_t(p+\phi, k-1) \right\} + \\
 & \alpha \left\{ \left(1 - \frac{1}{t} \right) n_t(k) + \frac{1}{t} n_t(k-1) + \delta_{k1} \right\},
 \end{aligned}$$

where $\phi = (1-p)(1-r)$ is the probability that only the old link is maintained in the DD step, and

$$F_t(x, k) = \sum_{j \geq k} \binom{j}{k} x^k (1-x)^{j-k} \frac{n_t(j)}{t}.$$

Note that $n_t(j) = 0$, if $j > t$ or $j < 0$. The recursion cannot in general be solved explicitly, but for a fixed choice of parameters it is easy to solve the recursion by

computational means. The master equation for DD+PA differs in the last term only;

$$\alpha \left\{ \left(1 - \frac{k}{\sum_j j n_t(j)} \right) n_t(k) + \frac{k-1}{\sum_j j n_t(j)} n_t(k-1) + \delta_{k1} \right\},$$

where further analysis of this expression is complicated because of the normalising sum.

It is natural to ask for properties of the expected degree sequence under DD+RA and DD+PA, e.g. whether the expected degree frequencies $f_t(k) = n_t(k)/t$, $k = 0, 1, \dots$, converge to a stationary distribution $f(k)$, $k = 0, 1, \dots$, as the network grows larger.⁵³

Theorem 2.1 (Pure DD, $\alpha = 0$). *We distinguish different scenarios:*

A *If $p < 1/2$, then there is a stationary distribution $\{f(k)\}_k$ as $t \rightarrow \infty$ (ergodic case).*

B *If*

$$\log(1 - \phi) + \log(p + \phi) + p < 0,$$

then the expected number $n_t(k)$ of nodes of degree k grows towards infinity for any $k \geq 0$, though there need not be a limiting distribution (recurrent case).

C *Finally, if*

$$\frac{1+p}{2+p} < (1-\phi)(p+\phi),$$

then there cannot be a limiting distribution and any infinitely large network contains a finite number of nodes of degree $k > 0$, but not necessarily of degree zero (transient case).

The proof can be found in Ref. 54; notably A implies B, but not *vice versa*.

Theorem 2.2 (DD+RA). *The theorem falls in two statements depending on α and p .*

A *If $(1 - \alpha)p < 1/2$, then there is a stationary distribution $\{f(k)\}_k$ as $t \rightarrow \infty$ (ergodic case).*

B *If $\alpha < 1$, then for any p , q and r the expected number $n_t(k)$ of nodes of degree k grows towards infinity for any $k \geq 0$, though there need not be a limiting distribution (recurrent case).*

The possibility to attach nodes randomly (RA) stabilises the network, such that there is no transient case for $\alpha < 1$. The mean, $M(1)$, of the degree distribution of a large network is finite exactly when $1 > 2(1 - \alpha)p$, and in that case

$$M(1) = \frac{2 - 2(1 - q)(1 - \alpha)}{1 - 2(1 - \alpha)p}.$$

When the mean exists, Theorems 2.1 and 2.2 tell us that there is a stationary distribution. For model DD+PA, this question has not been solved completely. The techniques applied in Ref. 54 are not directly transferable to model DD+PA, but it can be argued that Theorem 2.2B is true under the same circumstances (see also Theorem 2.4).

We now turn to the expected moments under models DD+RA and DD+PA. Let $M_t(i)$ be the i th descending moment of the degree, D_t , of a random node at step t ,

$$M_t(i) = \mathbb{E}[D_t(D_t - 1) \dots (D_t - i + 1)];$$

for example, $M_t(1)$ is the average node degree. The descending moments in DD+RA fulfil a simple recursion,

$$M_{t+1}(i) = \left(1 - \frac{\kappa(i)}{t+1}\right) M_t(i) + \frac{i\lambda(i)}{t+1} M_t(i-1),$$

where

$$\kappa(i) = 1 - (1 - \alpha)\{ip + (1 - \phi)^i + (p + \phi)^i - 1\}, \quad (2.1)$$

and

$$\lambda(i) = (1 - \alpha)q\{(1 - \phi)^{i-1} + (p + \phi)^{i-1}\} + (i - 1)(1 - \alpha)p + \alpha(1 + \delta_{i1}) \quad (2.2)$$

for $i \geq 1$ and $t \geq t_0$, and $M_t(0) = 1$ for all $t \geq t_0$.

Theorem 2.3 (DD+RA). *If $\kappa(i) > 0$ for $i \geq 1$, then $M_t(i)$, $t \geq t_0$, is converging with limit*

$$M(i) = \lim_{t \rightarrow \infty} M_t(i) = \frac{i! \prod_{j=1}^i \lambda(j)}{\prod_{j=1}^i \kappa(j)}.$$

If $\kappa(i) = 0$ and $\lambda(1) = 0$, then $\lim_{t \rightarrow \infty} M_t(i) = M_{t_0}(i)$. If $\kappa(i) < 0$, or if $\kappa(i) = 0$ and $\lambda(1) > 0$, then $M_t(i)$, $t \geq t_0$, increases beyond any bound.

Comparing model DD+RA to DD+PA, the first moments are identical, but higher moments differ.

Theorem 2.4 (DD+RA and DD+PA). *If $1 > 2(1 - \alpha)p$, then*

$$M(1) = \frac{2 - 2(1 - q)(1 - \alpha)}{1 - 2(1 - \alpha)p}.$$

If $1 = 2(1 - \alpha)p$ and $1 > (1 - q)(1 - \alpha)$, then $M_t(i) \propto \log(t)$, and if $1 < 2(1 - \alpha)p$, then $M_t(i)$, $t \geq t_0$, increases beyond any bound: $M_t(i) \propto t^{2(1 - \alpha)p - 1}$. Finally, in the remaining case $\alpha = 0$, $p = 1/2$ and $q = 0$, we have $M_t(1) = M_{t_0}(1)$ for all $t \geq t_0$.

It follows from Theorem 2.4 that if $\alpha = q = 0$ and $p < 1/2$, then $M(1) = 0$, so that the vast majority of nodes are of degree zero in a large network. Otherwise, at least a fraction $\alpha + (1 - \alpha)q$ of nodes has non-zero degree. From a biological perspective, nodes of degree zero represent non-functional genes. We neglect the

possibility for non-functional genes to revert to functional genes, by removing a node if its degree is zero when created. In practice, $q/t \approx 0$, so that this procedure is essentially equal to discarding the nodes of degree zero only *after* the network has been fully generated; in this latter situation, Theorems 2.1 and 2.2 remain valid as long as $\alpha > 0$ or $q > 0$.

Likewise, we can derive properties of the size of the interactome, i.e. the sum of all edges in the network, $I_t = tM_t(1)/2$ from Theorem 2.4. Notably, I_t attains a non-vanishing proportion of all possible edges $\binom{t}{2}$ only in the case where $p = 1$ and $\alpha = 0$.

2.4. Inferring Evolutionary Dynamics in Terms of Mixture Models of Network Growth

We have seen that it is very difficult to quantify the dynamics and modes of network evolution from PIN datasets analytically, and now turn to simulation-based tools. Adhering to an analysis that explicitly conditions on well-defined, clear models of network evolution, warrants ‘a meaningful comparison between the consequences of basic assumptions and the empirical facts’.⁵⁵ In this context, the Bayesian framework is our preferred method of statistical reasoning,⁵⁶ rather than optimisation or machine learning routines which often take a more implicit modelling approach. In Bayesian inference, the aim is to estimate the posterior density $p(\theta|\mathcal{G}_{\text{Obs}})$ of θ , given the observed network $\mathcal{G}_{\text{Obs}}$ under a given model, for example DD+PA. Bayes’ theorem relates $p(\theta|\mathcal{G}_{\text{Obs}})$ to the likelihood $L(\theta;\mathcal{G}_t) := \text{Prob}(\mathcal{G}_{\text{Obs}}|\theta)$ and the prior $p(\theta)$ by

$$p(\theta|\mathcal{G}_{\text{Obs}}) \propto L(\theta;\mathcal{G}_t)p(\theta). \quad (2.3)$$

In the absence of substantial prior information on the parameter values, we here use a uniform prior. In principle, this allows us to estimate the parameters of the model, and, provided the model is supported by the data, to test hypotheses about the network and the evolution of the interactome. For example, by comparing analyses from different species we might learn about the relative importance of different biological processes in the species and whether they evolve under similar constraints.

However, calculating the likelihood of a network under the evolutionary models of Sec. 2.3.2 has turned out to be a non-trivial task that requires advanced statistical tools and has only been accomplished for small and/or sparse biological networks.^{16,57} Here, we explain and develop these tools; we concentrate on the models DD+RA and DD+PA, though the presented techniques are applicable to a wide range of models of interactome evolution.

2.4.1. The likelihood of PIN data under DD+RA or DD+PA

Under the relatively complex models DD+RA or DD+PA, we are interested in calculating the likelihood $L(\theta; \mathcal{G}_t)$ of an observed network $\mathcal{G}_t$ for any $\theta = (\alpha, p, q, r)$. A sequence of events with graph rearrangements leading to a graph $\mathcal{G}_t$ is called a history of $\mathcal{G}_t$; i.e. the history is the sequence $\mathcal{H}_t = (\mathcal{G}_s, \mathcal{G}_2, \dots, \mathcal{G}_t)$, where $\mathcal{G}_s$ is the initial graph. Importantly, the joint likelihood of a graph and its history $L(\theta; \mathcal{G}_t, \mathcal{H}_t)$ is straightforward to calculate from the transition kernel of the models of network growth, whereas $L(\theta; \mathcal{G}_t)$ in principle requires summation over all possible histories. Formally, consider a graph $\mathcal{G}_t$ and denote the graph in which node v and all links to it are removed with $\delta(\mathcal{G}_t, v)$. A node v in $\mathcal{G}_t$ is said to be *removable* if $\mathcal{G}_t$ can be created by copying a node in $\delta(\mathcal{G}_t, v)$. If $\mathcal{G}_t$ contains removable nodes, it is said to be *reducible*, otherwise $\mathcal{G}_t$ is *irreducible*. Let $\mathcal{R}(\mathcal{G}_t)$ be the set of removable nodes. The likelihood can be written recursively

$$L(\theta, \mathcal{G}_t) = \frac{1}{t} \sum_{v \in \mathcal{R}(\mathcal{G}_t)} \omega_\theta(\mathcal{G}_t, v) L(\theta, \delta(\mathcal{G}_t, v)), \quad (2.4)$$

where $\omega_\theta(\mathcal{G}_t, v) = \text{Prob}(\mathcal{G}_t | \delta(\mathcal{G}_s, v), \theta)$.⁵⁷ The factor $1/t$ is the probability that v is the last added node, and the boundary condition for the recursion is $L(\theta; \mathcal{G}_s)$. For two histories $\mathcal{H}_t^1$ and $\mathcal{H}_t^2$ of a graph $\mathcal{G}_t$ starting from irreducible initial graphs $\mathcal{G}_s^1$ and $\mathcal{G}_s^2$, respectively, one can ask how different $\mathcal{G}_s^1$ and $\mathcal{G}_s^2$ can be. Surprisingly, the two graphs must be isomorphic to each other;⁵⁷ note that this statement is trivial when all nodes are removable, because we always end up with a graph consisting of one node. Therefore, we may put $L(\theta; \mathcal{G}_s) = 1$. If we could end up with non-isomorphic graphs (potentially with different number of nodes), then a (biologically non-trivial) prior distribution would be required for the initial graph in Eqn. (2.4).

Importantly, any network topology may be reproduced under models DD+RA and DD+PA.^{16,57} In particular, this property arises solely from the DD component as long as r does not equal zero or one, so that any (mixture) model including DD under the same conditions may explain the topology of real PIN datasets (of course, with different probabilities). In this respect, models DD+RA and DD+PA are more realistic than the models in Refs. 14,46,57, thus justifying their increased complexity.

Even though Eqn. (2.4) in principle provides the means to compute the likelihood, the method is computationally too intensive even for moderately sized PIN datasets $\mathcal{G}_{\text{obs}}$ under most mixture models of network growth. To see this for DD+PA or DD+RA, note that for most parameter values the set of removable nodes consists of all nodes in the network, $\mathcal{R}(\mathcal{G}_t) = \mathcal{V}_t$. This implies that any order of adding the nodes to the network is a history of the network, and consequently there are $t!$ different histories. Even if we keep a list of already calculated likelihoods, the number of recursive calls in Eqn. (2.4) is still immense. More importantly, Eqn. (2.4) is not well-suited to account for the following developments.

2.4.2. Simple methods to account for incomplete datasets

The fact that topological properties of incomplete PIN datasets may be biased to those of the (unknown) interactome,⁵⁸ necessitates a coherent account of the missing data. Incompleteness can be modelled by choosing randomly a subnet of a certain size from the full network; among others,⁴¹ two approaches are:^{59,60}

S1 A node is included in the subnet with probability $0 < \psi < 1$

S2 A node is pre-selected with probability $0 < \psi < 1$. If its degree among pre-selected nodes is not zero, then it is included in the subnet.

The full genome size t is still not known precisely for most organisms; an estimate might be obtained from the consensus number of open reading frames (ORFs), see Table 2.1. Although it is in principle possible to account for uncertainty in t within our Bayesian perspective, we here assume t is fixed. It then follows under S1 that the sampling fraction can be estimated by $\hat{\psi} = V/t$. Under S2, the estimate cannot be calculated analytically (unless the experimenter reveals the number of proteins with observed degree zero), but must be estimated together with θ . In practice, $\hat{\psi} = V/t$ is a reasonable estimate under both sampling schemes.

The qualitative effect of sampling on network quantities has been studied to some extent.⁶⁰ Let D_t^* denote the degree of a node drawn according to S1. The variables D_t and D_t^* are related through $D_t^* \sim Bi(D_t, \psi)$, i.e. given $D_t = d$, D_t^* is drawn from the binomial distribution $Bi(d, \psi)$. It follows that the factorial moments, $M_t^{S1}(i)$, $i \geq 1$, in the subnet under S1 take the form⁵⁹

$$M_t^{S1}(i) = \mathbb{E}[(D_t^*)_{[i]}] = \psi^i \mathbb{E}[(D_t)_{[i]}] = \psi^i M_t(i).$$

Under S2, the moments take the form

$$M_t^{S2}(i) = \frac{\mathbb{E}[(D_t^*)_{[i]}]}{P(D_t^* > 0)} = \frac{\psi^i \mathbb{E}[(D_t)_{[i]}]}{P(D_t^* > 0)} = \frac{\psi^i M_t(i)}{P(D_t^* > 0)}.$$

Whereas the moments under S1 are easily derived from the expressions in Eqns. (2.1) and (2.2), the moments under S2 are not easily evaluated unless we know the degree sequence. We have $P(D_t^* > 0) = 1 - \mathbb{E}[(1 - p)^{D_t}]$. Remarkably, the relative moments are the same under the two sampling schemes,

$$\frac{\psi M_t(i+1)}{M_t(i)} = \frac{M_t^{S1}(i+1)}{M_t^{S1}(i)} = \frac{M_t^{S2}(i+1)}{M_t^{S2}(i)}.$$

When computing the likelihood recursively, it is not possible to account for incompleteness. This motivated us, together with the fact that computational considerations limit the range of entertainable models, to devise alternative, more approximate methods than Eqn. (2.4). Importantly, these approaches also afford to incorporate noise and sampling bias into the computational analysis, aspects of network inference which are difficult to study qualitatively.

Table 2.2. Summary Statistics.

Order	The number of nodes in a network
Size	The number of edges in a network
Degree	The number of edges associated with a node
ND	Degree sequence, $p(D = k)$, the percentage of nodes with degree $k = 0, 1, \dots$ in a network
$\overline{\text{ND}}$	Average node degree, the mean degree of a network
CC	Average cluster coefficient, mean probability that two neighbours of a node are themselves neighbours
Distance	The minimum number of edges that have to be visited to reach a node j from node i
CONN	Relative log connectivity distribution, $\log(p(k_1, k_2)\overline{\text{ND}}^2)/(k_1 p(k_1)k_2 p(k_2))$, the depletion or enrichment of edges ending in nodes of degree k_1, k_2 relative to the uncorrelated network with the same ND ¹⁰
WR	Within-reach distribution, $p(\text{WR} \leq k)$, the mean probability of how many nodes are reached from one node within distance $k = 1, 2, \dots$ in the network ¹⁶
DIA	Diameter, the longest minimum path among pairs of nodes in a connected component of the network
FRAG	Fragmentation, the percentage of nodes not in the largest connected component

2.4.3. Approximating the likelihood with many summaries

Instead of calculating the likelihood of the full observed network, we may reduce the network to a set of summary statistics $\mathbb{S} = (S_1, \dots, S_K)$, and consider $L(\mathbb{S}(\mathcal{G}_{\text{Obs}}); \theta, \psi)$ rather than $L(\mathcal{G}_{\text{Obs}}; \theta, \psi)$ for inference. Typically, $\mathbb{S}$ is of lower dimension than $\mathcal{G}$, such that complex models of network evolution may be amenable for statistical analysis. If $\mathbb{S}$ is sufficient for a model parameter θ , then the posterior of θ given $\mathcal{G}_{\text{Obs}}$ is the same as the posterior of θ given $\mathbb{S}(\mathcal{G}_{\text{Obs}})$. For example, consider the parameters θ and ψ under the ER graph. Since the probability of a graph,

$$\binom{M}{|\mathcal{E}_t|} \theta^{|\mathcal{E}_t|} (1 - \theta)^{M - |\mathcal{E}_t|},$$

where $M = \binom{t}{2}$, depends on the link probability θ only through the number of links $|\mathcal{E}_t|$, it is a sufficient statistic for θ . Accounting for incompleteness with S1, the probability becomes

$$\binom{M_{\text{Obs}}}{|\mathcal{E}_{\text{Obs}}|} (\psi\theta)^{|\mathcal{E}_{\text{Obs}}|} (1 - \psi\theta)^{M_{\text{Obs}} - |\mathcal{E}_{\text{Obs}}|},$$

where $M_{\text{Obs}} = \binom{V_{\text{Obs}}}{2}$. Consequently, $|\mathcal{E}_t|$ is now a sufficient statistic for the product $\psi\theta$; unless we treat ψ as known (which we generally do), we cannot separate inference on ψ and θ .

For complex models of network growth, low-dimensional summary statistics are unknown, and $p(\theta|\mathbb{S}(\mathcal{G}_{\text{Obs}}))$ is taken as an *approximation* of $p(\theta|\mathcal{G}_{\text{Obs}})$; approximation quality then has to be analysed separately and generally depends on $\mathbb{S}$. The set of summaries could be the degree sequence alone,⁶¹ the lowest degree moments or some other characteristics of the network; see Table 2.2 for those we apply here.

2.4.4. Approximate Bayesian computation

Likelihood-free inference (LFI) confers computational tractability by comparing simulated data $\mathcal{G}$ to the observed data $\mathcal{G}_{\text{Obs}}$ instead of calculating the likelihood directly. Approximate Bayesian computation (ABC), reviewed in Ref. 62, is a powerful implementation of LFI. It may be interpreted as approximating the likelihood with

$$L_K(\theta; \mathcal{G}_{\text{Obs}}) = \int K(\mathcal{G}_{\text{Obs}}|\mathcal{G})p(\mathcal{G}|\theta)d\mathcal{G}, \quad (2.5)$$

where $K(\mathcal{G}_{\text{Obs}}|\mathcal{G})$ is a suitable, weighted measure of the proximity of the simulated to the observed data; the approximate posterior follows in analogy to Eqn. (2.3),

$$p_K(\theta|\mathcal{G}_{\text{Obs}}) \propto L_K(\theta; \mathcal{G}_{\text{Obs}})p(\theta). \quad (2.6)$$

In practice, numerical estimates $\tilde{p}_K(\theta|\mathcal{G}_{\text{Obs}})$ of Eqn. (2.6) may be obtained with a variety of Monte Carlo strategies.⁶³ All methods of ABC are based around the particularly simple kernel $K_{\text{ABC}}(\mathcal{G}_{\text{Obs}}|\mathcal{G}) = \mathbf{1}\{d(\mathbb{S}(\mathcal{G}), \mathbb{S}(\mathcal{G}_{\text{Obs}})) \leq h\}$, which compares $\mathcal{G}$ to $\mathcal{G}_{\text{Obs}}$ in terms of a set of (computationally tractable) summaries $\mathbb{S} = (S_1, \dots, S_k, \dots, S_K)$ under a distance function d and fixed, non-negative mismatch threshold h . In practice, h is chosen as small as possible, implicitly assuming that the underlying model is correct.

For network data, embedding LFI into Markov Chain Monte Carlo (MCMC) is particularly attractive.¹⁶ The algorithm proceeds as follows:

- MC1** Compute the observed summaries $\mathbb{S}(\mathcal{G}_{\text{Obs}})$ and start at some initial value θ
- MC2** If now at θ , propose a move to θ' according to a proposal density $q(\theta \rightarrow \theta')$; here we take a Gaussian, centred at θ with diagonal covariance matrix Σ , restricted to the interval $[0, 1]$
- MC3** Given θ' , grow a dataset to the estimated genome size reported in Table 2.1. Take a random subnet $\mathcal{G}'$ that matches the order of the observed PIN dataset, and compute $\mathbb{S}(\mathcal{G}')$
- MC4** Accept θ' with probability

$$\min \left(1, \frac{p(\theta')q(\theta' \rightarrow \theta)}{p(\theta)q(\theta \rightarrow \theta')} \mathbf{1}\{d(\mathbb{S}(\mathcal{G}_{\text{Obs}}), \mathbb{S}(\mathcal{G}')) \leq h\} \right),$$

and otherwise stay at θ , then return to MC2. Here, $\mathbf{1}$ denotes the indicator function, $h = (h_1, \dots, h_k)$ is a threshold vector and $d = (d_1, \dots, d_k)$ a function such that d_j is a distance on S_j for all j . The notation $d(\mathbb{S}(\mathcal{G}_{\text{Obs}}), \mathbb{S}(\mathcal{G}')) \leq h$ means that the inequality is fulfilled for all j .

This algorithm is guaranteed to eventually generate a series of correlated samples from

$$p(\theta|d(\mathbb{S}(\mathcal{G}_{\text{Obs}}), \mathbb{S}(\mathcal{G})) \leq h). \quad (2.7)$$

When h_j , $j = 1, \dots, k$ approach zero, the posterior density Eqn. (2.7) approaches $p(\theta|\mathbb{S}(\mathcal{G}_{\text{Obs}}))$. However, the above algorithm will then often fail or become inefficient unless the observed data is frequently reproduced under the model, because the acceptance probability in MC4 also approaches zero. On the other hand, if h_j , $j = 1, \dots, k$ are large, the above algorithm becomes more efficient but Eqn. (2.7) approaches the prior of θ , $p(\theta)$. Choosing appropriate values of h_j is a technical issue that must be addressed carefully. Even with a sensible choice of h , convergence of algorithm MC1–MC4 is not straightforward and requires a number of technical modifications outlined in Ref. 16.

Choosing appropriate summaries and distance functions is crucial to ensure the approximation quality of Eqn. (2.7) to the likelihood in the absence of a general approximation theory.⁶² For consistent and reliable parameter inference on PINs, we have demonstrated¹⁶ that the observed data is best described by a comprehensive set of summaries under a strict approximation criterion that requires *separate* h_j for each summary S_j . Figure 2.4 illustrates the difference between using a single summary statistics and a set of summaries. In passing, we note that computational methods that target Eqn. (2.6) are required not to suffer from the inclusion of many summaries, and MCMC appears as a viable, computational device. In an extensive consistency analysis, we have determined suitable, comprehensive sets of summaries, one of which is $\mathbb{S}^* = \text{WR, DIA, } \overline{\text{ND}}, \text{CC, FRAG}$.¹⁶ In addition, we found that the degree sequence alone and motif counts have very limited value in estimating the model parameters.¹⁶ Good summaries are thus not necessarily those that are amenable to a rigorous mathematical analysis as in Sec. 2.3.3; this highlights the importance of simulation-based methods, but also warns that our intuition, in the guise of analytical formulae, might be limited to *relatively* uninformative aspects of biological networks, particularly when they are not considered in context.

2.4.5. *Evolutionary analysis of the PIN topologies of T. pallidum, H. pylori and P. falciparum*

We illustrate the ability of LFI to provide quantitative, reliable estimates of broad evolutionary parameters under model DD+PA. This model was designed to quantify whether the likelihood of gene duplication plays a larger role in network evolution of eukaryotes than prokaryotes. We consider here the three small PIN datasets of the prokaryotes *T. pallidum*, *H. pylori*, and the eukaryote *P. falciparum*. The fact that a reliable, consistent analysis requires the combination of several summaries that capture global aspects of the networks, renders an implementation targeting, for example, the *S. cerevisiae* PIN dataset, computationally challenging.

We successfully applied a technical variant of algorithm MC1–MC4 to all three PIN datasets based on the set of summaries $\mathbb{S}^*$ under model DD+PA; the mismatch thresholds were determined in preliminary test runs to ensure approximation and mixing quality of the algorithm, see Ref. 16. Figure 2.5 displays the one-dimensional

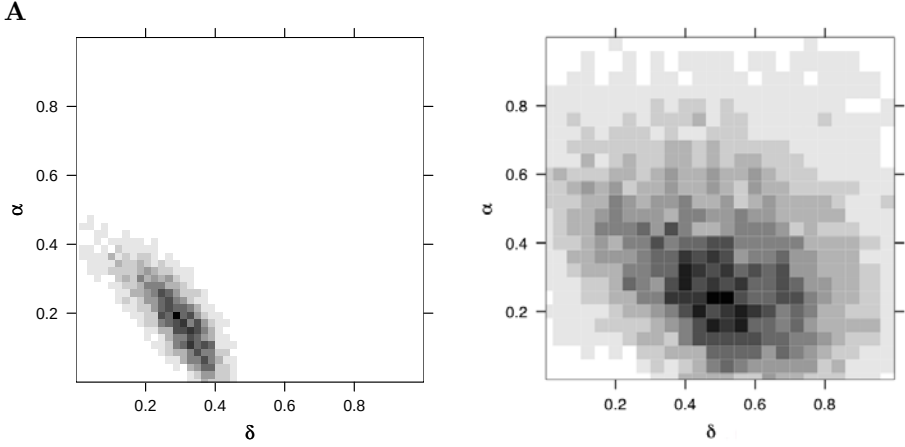


Fig. 2.4. For the *H. pylori* PIN data, comparison of inference using one versus four summary statistics. (A) 2D-histogram of the posterior parameters (α, δ) , with $\delta = (1 - p)/(1 + p)$, obtained with $\mathbb{S}^*$. Posterior mass clearly centres on a tight cloud in the parameter space. (B) The same but using only ND. The regions of highest posterior density using ND are inconsistent with those using $\mathbb{S}^*$; see Ref. 16 for details.

Table 2.3. Estimated evolutionary dynamics of *T. pallidum*, *H. pylori* and *P. falciparum*, with $\delta = (1 - p)/(1 + p)$.

Species	δ	p	q	α
<i>T. pallidum</i>	0.34 (0.13,0.49)	0.49 (0.34,0.77)	0.32 (0.08,0.67)	0.28 (0.05,0.55)
<i>H. pylori</i>	0.28 (0.14,0.39)	0.56 (0.44,0.75)	0.05 (0.01,0.10)	0.22 (0.08,0.36)
<i>P. falciparum</i>	0.32 (0.26,0.37)	0.52 (0.46,0.59)	0.05 (0.00,0.09)	0.07 (0.02,0.13)

MCMC trace plots of $\alpha \in (0, 1)$ for the *H. pylori* and *P. falciparum* datasets, indicating good convergence; similar results are obtained for all other model parameters across all organisms. Table 2.3 lists the 80% credible intervals (i.e. the inner range of values of a random variable that attains 80% posterior mass) of θ under model DD+PA for all PIN datasets. Notably, the DD component obtained considerably, but not significantly, less posterior weight for the two prokaryotic PIN datasets than for the eukaryote. This is in accordance with current beliefs that other processes than gene duplication (DD) play an important role in the evolution of prokaryotic networks.¹⁹

The interpretation of the approximate posterior densities must be considered within the limits of the model, the data and the approximative nature of the inference method. For example, sampling bias of PIN datasets may not be adequately addressed by taking random subsamples of simulated networks that are grown to the estimated number of open reading frames; see also Sec. 2.2 and Sec. 2.4.4. Re-

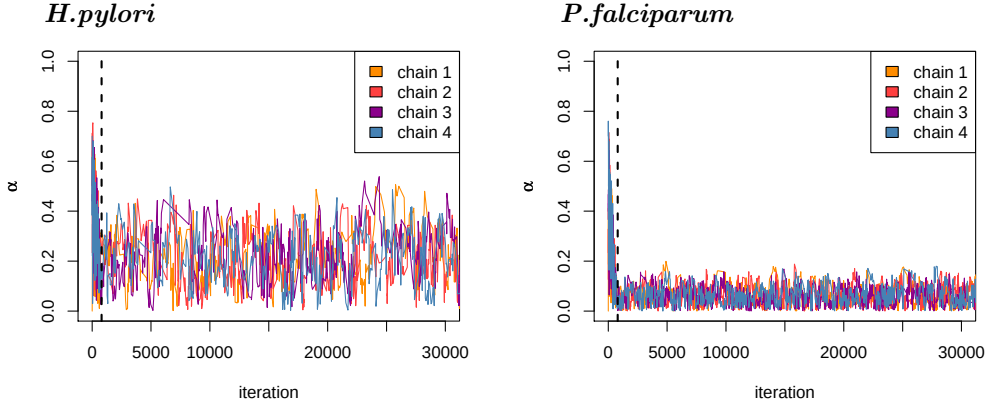


Fig. 2.5. Traceplots of $\alpha \in (0,1)$ from the MCMC output for the *H. pylori* and *P. falciparum* datasets. Four MCMC chains were run for 75,000 iterations (the first 30,000 are shown here) according to MC1–MC4 based on S^* from overdispersed initial values. The chains converge quickly within the burn-in period (iteration 800, vertical dashed line); thereafter moves are taken to represent samples from the posterior.

assuringly, the credibility intervals of *P. falciparum* overlap nicely with parameter estimates obtained from sequence data of *S. cerevisiae*, where a mean divergence probability ($\delta = (1-p)/(1+p)$) of around 35%–42% and a mean attachment probability (q) of around 1%–2% within the first 25Myr after a duplication event have been reported.⁹ Further, we cannot explain the marked difference in posterior estimates of q between *T. pallidum* and *H. pylori*. This suggests that, alternatively, differences in the experimental protocol to obtaining high-throughput PIN data may confound our evolutionary analysis of network topologies from different domains.

We note that the values of p , q and α reported in Table 2.3 suggest that a stationary degree distribution does not exist for *H. pylori* and *P. falciparum*, whereas it may for *T. pallidum* (see Theorem 2.2). Under the assumption that the model is correct, this indicates that key characteristics of a network, such as degree distribution, are not time-invariant as evolution modifies the network.

2.4.6. The size of the interactome

Aspects of the complete, unobserved interactome are easily predicted from the noisy and incomplete observed PIN data, once MCMC output is available. Here, we briefly discuss the interactome by means of its posterior predictive distribution. The posterior predictive distribution for *H. pylori* has a mode of 5,636 and 80% credibility interval (2,915; 8,536), whereas for *P. falciparum* the mode is 43,835 and the credibility interval is (18,689; 84,205). These compare with estimates obtained by other means; e.g. Ref. 64 reports 6,082 and 45,940 for *H. pylori* and *P. falciparum*, respectively, and using the method in Ref. 65 we obtain 5,412 and 45,868, respectively.

2.5. Conclusion

We have showed that it is possible to draw quantitative, evolutionary inferences from large-scale, incomplete network data with extensive computer simulations that explicitly condition on well-defined models of network growth. Using a likelihood-free approach that relies on comparing summaries of real network data to simulated PINs, we were able to study more complex models of network evolution more confidently than had been previously possible. Crucially, we found that these complex models are more realistic than previous models, in that the topology of real networks may be fully explained, at least in a qualitative sense. These mixture models of network growth are hard to analyse rigorously; only some asymptotic properties of particular, amenable aspects of networks (generated under these models) could be derived. Importantly, the set of summaries that proved most useful in our simulation-based analysis did not include any of the analytically tractable summaries. Thus, in the absence of a thorough understanding of the workings of the models, we recommend careful interpretation of the achieved results.

Here, we have focused on a particular model of network evolution, DD+PA. Naturally, our interpretations of the estimated model parameters are conditional, not only on the quality of the PIN datasets, but also on the particular model under consideration, the employed sampling scheme, as well as the choice of data used to inform the presented analyses. We have recently generalised the presented framework of likelihood-free inference to account more explicitly for the underlying model.⁶⁶ Perhaps along these lines, more work may provide a fuller statistical analysis of interactome evolution.

Acknowledgements

Carsten Wiuf is supported by the Danish Cancer Society and the Danish Research Councils. Oliver Ratmann is supported by the Wellcome Trust, UK.

Appendix A. Proofs of Theorems.

The descending moments in DD-RA fulfil a simple recursion,

$$M_{t+1}(i) = \left(1 - \frac{\kappa(i)}{t+1}\right)M_t(i) + \frac{i\lambda(i)}{t+1}M_t(i-1), \quad (\text{A.1})$$

where

$$\kappa(i) = 1 - (1 - \alpha)\{ip + (1 - \phi)^i + (p + \phi)^i - 1\}, \quad (\text{A.2})$$

and

$$\lambda(i) = (1 - \alpha)q\{(1 - \phi)^{i-1} + (p + \phi)^{i-1}\} + (i - 1)(1 - \alpha)p + \alpha(1 + \delta_{i1}) \quad (\text{A.3})$$

for $i \geq 1$ and $t \geq t_0$, and $M_t(0) = 1$ for all $t \geq t_0$.

An argument for Eqn. (A.1) can be obtained by multiplying the master equation by $k(k-1)\dots(k-i+1)$ and summing over all k .

Lemma 2.1. *Assume $\kappa(1) > 0$. The moments $M_t(1)$, $t \geq t_0$, fulfil*

$$M_{t+1}(1) > M_t(1) \Leftrightarrow \frac{\lambda(1)}{\kappa(1)} > M_t(1). \quad (\text{A.4})$$

If the statement holds for $t = t_0$, it holds for all $t \geq t_0$, and as a consequence $M_t(1)$, $t \geq t_0$, is converging.

Proof. [Proof of Lemma 2.1] It follows from Eqn. (A.1) that

$$M_{t+1}(1) = \left(1 - \frac{\kappa(1)}{t+1}\right) M_t(1) + \frac{\lambda(1)}{t+1} > M_t(1),$$

if and only if Eqn. (A.4) is true. Assume the statement is true for all t in $s \geq t \geq t_0$. Then

$$M_{s+1}(1) = \left(1 - \frac{\kappa(1)}{s+1}\right) M_s(1) + \frac{\lambda(1)}{s+1} <$$

$$\left(1 - \frac{\kappa(1)}{s+1}\right) \frac{\lambda(1)}{\kappa(1)} + \frac{\lambda(1)}{s+1} = \frac{\lambda(1)}{\kappa(1)}$$

and the statement is true for $s+1$. It follows that $M_t(1)$, $t \geq t_0$, is converging, either because the inequality $M_{t+1}(1) > M_t(1)$ is fulfilled or the reverse inequality. The proof of the lemma is completed. $\square$

Lemma 2.2. *Assume $\kappa(i) > 0$ for $i \geq 2$. The moments $M_t(i)$, $t \geq t_0$, fulfil*

$$M_{t+1}(i) > M_t(i) \Leftrightarrow \frac{i\lambda(i)}{\kappa(i)} M_t(i-1) > M_t(i). \quad (\text{A.5})$$

If $M_{t+1}(i) > M_t(i)$, then also

$$\frac{i\lambda(i)}{\kappa(i)} M_t(i-1) > M_{t+1}(i), \quad (\text{A.6})$$

and likewise with $>$ replaced by $\leq$.

Proof. [Proof of Lemma 2.2] It follows from Eqn. (A.1) that

$$M_{t+1}(i) = \left(1 - \frac{\kappa(i)}{t+1}\right) M_t(i) + \frac{i\lambda(i)}{t+1} M_t(i-1) > M_t(i),$$

if and only if Eqn. (A.5) is true. Assume $M_{t+1}(i) > M_t(i)$. Then

$$M_{t+1}(i) = \left(1 - \frac{\kappa(i)}{t+1}\right) M_t(i) + \frac{i\lambda(i)}{t+1} M_t(i-1) <$$

$$\left(1 - \frac{\kappa(i)}{t+1}\right) \frac{i\lambda(i)}{\kappa(i)} M_t(i-1) + \frac{i\lambda(i)}{t+1} M_t(i-1) = \frac{i\lambda(i)}{\kappa(i)} M_t(i-1),$$

which is the inequality to be proven. The proof of the lemma is completed. $\square$

Lemma 2.3. *There exists $J \geq 1$ (potentially ∞), such that $\kappa(i) > 0$ for all $1 \leq i < J$, $\kappa(J) \leq 0$ and $\kappa(i) < 0$ for $i > J$.*

Proof. [Proof of Lemma 2.3] Define $A(x) = 1 - (1 - \alpha)\{xp + (1 - \phi)^x + (p + \phi)^x - 1\}$ for $x \geq 0$, and note that $A(i) = \kappa(i)$. By differentiation, $A''(x) \leq 0$ for all $x \geq 0$ and $A(x)$ is concave. Let J be the first integer such that $\kappa(J) \leq 0$ (if it exists). If $J > 1$, then $\kappa(J - 1) > 0$ and the result follows from concavity. For $J = 1$, there are several cases: 1) If $\kappa(1) < 0$, then it follows from concavity since $A(0) = \alpha \geq 0$. 2) If $\kappa(1) \leq 0$ and $\alpha > 0$, then it follows from concavity since $A(0) = \alpha > 0$. 3) If $\kappa(1) = 0$ and $\alpha = 0$, then $p = 1/2$ and consequently $\kappa(2) \leq -1/8$. By concavity, it follows that $\kappa(i) < 0$ for $i > 2$. The proof is completed. $\square$

Lemma 2.4. *Assume $\lambda(j) = 0$ for some $j > 1$. Then $\lambda(i) = 0$ for all $i \geq 1$, and consequently $\kappa(i) > 0$ for all $i \geq 1$. (Note that $\lambda(1) = 0$ does not imply that $\lambda(i) = 0$ for any $i > 1$.)*

Proof. [Proof of Lemma 2.4] Assume $\lambda(j) = 0$ for some $j > 1$. From Eqn. (A.3) with $i = j > 1$, it follows that $\alpha = p = q = 0$ and consequently $\lambda(i) = 0$ for all $i \geq 1$. From Eqn. (A.2), it follows that $\kappa(i) > 0$ for all $i \geq 1$. $\square$

Lemma 2.5. *Assume $\kappa(i) > 0$ for some $i \geq 1$ and $\lambda(1) = 0$. Then there exists a constant $C_i > 0$ such that*

$$M_t(i) \leq C_i t^{-a_i}, \quad (\text{A.7})$$

where a_i is any positive number such that $a_i < \min\{\kappa(j) | 1 \leq j \leq i\}$. Note that C_i is specific to the particular i , while a_i needs to be chosen relatively to all $\kappa(j)$, $j \leq i$.

Proof. [Proof of Lemma 2.5] First note that for $\kappa(j) > 0$ there exist constants $d_j > 0$ and $D_j > 0$, such that

$$D_j t^{-\kappa(j)} \leq \prod_{s=t_0}^t \left(1 - \frac{\kappa(j)}{s}\right) \leq D_j t^{-\kappa(j)} \quad (\text{A.8})$$

for all $t \geq t_0$. The proof of the lemma is by induction in i . For $i = 1$ (with $\lambda(1) = 0$),

$$M_{t+1}(1) = \left(1 - \frac{\kappa(1)}{t+1}\right) M_t(1).$$

Consequently

$$M_{t+1}(1) = \prod_{s=t_0+1}^{t+1} \left(1 - \frac{\kappa(1)}{s}\right) M_{t_0}(1),$$

and the result follows from Eqn. (A.8) with $a_1 < \kappa(1)$ (in fact, equality holds in this case). Next, assume it is true for $j \leq i - 1$ and consider Eqn. (A.1) for $M_t(i)$:

$$M_{t+1}(i) = \left(1 - \frac{\kappa(i)}{t+1}\right) M_t(i) + \frac{i\lambda(i)}{t+1} M_t(i-1).$$

It follows from Lemma 2.3 that $\kappa(j) > 0$ for all $1 \leq j \leq i$; hence also that

$$M_t(i-1) \leq C_{i-1} t^{-a_{i-1}}$$

for $a_{i-1} < \min\{\kappa(j) | 1 \leq j \leq i-1\}$ and $t \geq t_0$. Then

$$M_{t+1}(i) \leq \left(1 - \frac{\kappa(i)}{t+1}\right) M_t(i) + \frac{i\lambda(i)}{t+1} C_{i-1} t^{-a_{i-1}}.$$

By repeated application of Eqn. (A.1),

$$M_{t+1}(i) \leq \sum_{s=t_0+1}^{t+1} \frac{i\lambda(i)}{s} \frac{C_{i-1}}{(s-1)^{a_{i-1}}} \prod_{u=s+1}^{t+1} \left(1 - \frac{\kappa(i)}{u}\right),$$

and by manipulating the terms using Eqn. (A.8),

$$M_{t+1}(i) \leq \frac{C_i}{(t+1)^{a_i}} \sum_{s=t_0+1}^{t+1} \frac{1}{s},$$

where $a_i < \min\{\kappa(j) | 1 \leq j \leq i\}$. The constant C_i depends on the various constants in the sum as well as on d_i and D_i . Note that $\log(t)/t^\epsilon \rightarrow 0$ as $t \rightarrow \infty$ for any $\epsilon > 0$; hence

$$M_{t+1}(i) \leq \frac{C_i}{(t+1)^{a_i}}$$

for $a_i < \min\{\kappa(j) | 1 \leq j \leq i\}$, and the lemma is proved. $\square$

Theorem 2.3. If $\kappa(i) > 0$ for $i \geq 1$, then $M_t(i)$, $t \geq t_0$, is converging with limit

$$M(i) = \lim_{t \rightarrow \infty} M_t(i) = \frac{i! \prod_{j=1}^i \lambda(j)}{\prod_{j=1}^i \kappa(j)}. \quad (\text{A.9})$$

If $\kappa(i) = 0$ and $\lambda(1) = 0$, then $\lim_{t \rightarrow \infty} M_t(i) = M_{t_0}(i)$. If $\kappa(i) < 0$, or if $\kappa(i) = 0$ and $\lambda(1) > 0$, then $M_t(i)$, $t \geq t_0$, increases beyond any bound.

Proof. [Proof of Theorem 2.3] The proof is by induction. Assume $i = 1$ and $\kappa(1) > 0$. It follows from Eqn. (A.4) that $M_t(1)$ is converging. We have

$$M_{t+1}(1) - M_t(1) = \frac{1}{t+1} [\lambda(1) - \kappa(1)M_t(1)]. \quad (\text{A.10})$$

If $\lim_{t \rightarrow \infty} M_t(1) \neq \lambda(1)/\kappa(1)$, then it follows from Eqn. (A.10) that $M_t(1)$ is increasing or decreasing without bound, contradicting that $M_t(1)$ is converging. Hence

$$\lim_{t \rightarrow \infty} M_t(1) = \frac{\lambda(1)}{\kappa(1)}.$$

If $\kappa(i) > 0$, then $\kappa(j) > 0$ for all $1 \leq j \leq i$ according to Lemma 2.3. Assume the theorem is true for $i-1$, i.e. that $M_t(j)$, $t \geq t_0$, is converging for all $i-1 \geq j \geq 1$ with limit given by Eqn. (A.9).

First we will prove that $M_t(i)$, $t \geq t_0$, is converging. Define S_ϵ such that

$$|M_t(i-1) - K_{i-1}| \leq \epsilon$$

for $t \geq S_\epsilon$ and $\epsilon > 0$, where K_{i-1} denotes the limit of $M_t(i-1)$. Further define T_ϵ by

$$T_\epsilon = \min\{t > S_\epsilon \mid M_{t-1}(i) < M_t(i) \text{ and } M_t(i) \geq M_{t+1}(i)\}.$$

If $T_\epsilon = \infty$, then either $M_t(i)$ is increasing from a certain point $t \geq S_\epsilon^* > S_\epsilon$, or $M_t(i)$ is decreasing for all $t > S_\epsilon$. In the first case, it follows from Lemma 2.2 that

$$\frac{i\lambda(i)}{\kappa(i)}(K_{i-1} + \epsilon) > M_t(i)$$

for all $t \geq S_\epsilon^*$; hence $M_t(i)$, $t \geq S_\epsilon^*$, is increasing and bounded, thus also converging. In the latter case, it likewise follows from Lemma 2.2 that

$$\frac{i\lambda(i)}{\kappa(i)}(K_{i-1} - \epsilon) \leq M_t(i)$$

for all $t > S_\epsilon$; hence $M_t(i)$, $t \geq 1$, is converging. If $T_\epsilon < \infty$, then

$$\frac{i\lambda(i)}{\kappa(i)}(K_{i-1} - \epsilon) < M_t(i) < \frac{i\lambda(i)}{\kappa(i)}(K_{i-1} + \epsilon) \quad (\text{A.11})$$

for all $t \geq T_\epsilon$. The proof of this fact is by induction. First, Lemma 2.2 shows that $t = T_\epsilon$ fulfils Eqn. (A.11). Assume Eqn. (A.11) is fulfilled for $s \geq t \geq T_\epsilon$ for some s . Consider $t = s + 1$. Either $M_{s+1}(i) > M_s(i)$, or $M_{s+1}(i) \leq M_s(i)$. In the first case, Lemma 2.2 shows that $[i\lambda(i)/\kappa(i)](K_{i-1} + \epsilon) > M_{s+1}(i)$, and since $M_s(i)$ is bounded from below, so is $M_{s+1}(i)$. Hence Eqn. (A.11) is fulfilled for $t = s + 1$.

The latter case follows similarly. Hence for all $t \geq T_\epsilon$, Eqn. (A.11) is true. Since it holds for for any $\epsilon > 0$, $M_t(i)$, $t \geq 1$, is converging. The proof (by induction) that $M_t(i)$, $t \geq t_0$, is converging is completed.

The form of the limit also follows by induction. For $i = 1$ it is proven above. Assume the limit takes the form stated in the theorem for $i - 1$. Then it follows from the two inequalities in Eqn. (A.11) that the form is also correct for $i = 1$. The proof of the case $\kappa(i) > 0$ is completed.

If $\kappa(i) = \lambda(1) = 0$, $i > 1$, then it follows from Lemma 2.3 that $\kappa(j) > 0$ for all $1 \leq j \leq i - 1$. Hence it follows from Lemma 2.5 and Eqn. (A.1) that

$$M_t(i) \leq M_{t+1}(i) \leq M_t(i) + \frac{i\lambda(i)C_{i-1}}{t+1}t^{-a_{i-1}}.$$

Repeated iterations yield

$$M_{t_0}(i) \leq M_{t+1}(i) \leq M_{t_0}(i) + \sum_{s=t_0}^t \frac{i\lambda(i)C_{i-1}}{s+1}s^{-a_{i-1}}.$$

The sum is easily seen to converge towards zero; hence $\lim_{t \rightarrow \infty} M_t(i) = M_{t_0}(i)$.

If $\kappa(i) < 0$, then it follows from Eqn. (A.1) that $M_t(i)$ increases beyond any bound. If $\kappa(i) = 0$ and $\lambda(1) > 0$, then also $\lambda(i) > 0$ (Lemma 2.4) and it follows from Eqn. (A.1) that $M_t(i)$ increases towards infinity. $\square$

References

1. M. Monica, Genomes, phylogeny, and evolutionary systems biology, *Proceedings of the National Academy of Sciences*. **102**(suppl.1), 6630–6635, (2005).
2. J. S. Weitz, P. N. Benfey, and N. S. Wingreen, Evolution, interactions, and biological networks, *PLoS Biology*. **5**(1), (2007).
3. M. F. Oleksiak, G. A. Churchill, and D. L. Crawford, Variation in gene expression within and among natural populations, *Nat Genet*. **32**(2), 261–266, (2002).
4. A. P. P. Gasch, A. M. M. Moses, D. Y. Y. Chiang, H. B. B. Fraser, M. Berardini, and M. B. B. Eisen, Conservation and evolution of cis-regulatory systems in ascomycete fungi., *PLoS Biol*. **2**(12) (November, 2004).
5. A. Tanay, A. Regev, and R. Shamir, Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast, *Proceedings of the National Academy of Sciences of the United States of America*. **102**(20), 7203–7208, (2005).
6. L. Marino-Ramirez, I. K. Jordan, and D. Landsman, Multiple independent evolutionary solutions to core histone gene regulation, *Genome Biology*. **7**(12), R122, (2006).
7. E. H. Davidson and D. H. Erwin, Gene regulatory networks and the evolution of animal body plans, *Science*. **311**(5762), 796–800, (2006).
8. M. Lynch, The evolution of genetic networks by non-adaptive processes, *Nat Rev Genet*. **8**(10), 803–813, (2007).
9. A. Wagner, How the global structure of protein interaction networks evolves, *Proceedings: Biological Sciences*. **270**(1514), 457–466, (2003).
10. J. Berg, M. Lässig, and A. Wagner, Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications, *BMC Evol. Biol*. **4**, 51, (2004).
11. S. Wuchty, Evolution and topology in the yeast protein interaction network, *Genome Research*. **14**(7), 1310–1314, (2004).
12. P. Beltrao and L. Serrano, Specificity and evolvability in eukaryotic protein interaction networks., *PLoS Comput Biol*. **3**(2), e25, (2007).
13. K. Evlampiev and H. Isambert, Modeling protein network evolution under genome duplication and domain shuffling, *BMC Systems Biology*. **1**(1), 49, (2007).
14. K. Evlampiev and H. Isambert, Conservation and topology of protein interaction networks under duplication-divergence evolution, *Proceedings of the National Academy of Sciences*. **105**(29), 9863–9868, (2008).
15. C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, Evolution of the Protein Repertoire, *Science*. **300**(5626), 1701–1703, (2003).
16. O. Ratmann, O. Jørgensen, T. Hinkley, M. P. Stumpf, S. Richardson, and C. Wiuf, Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H.pylori* and *P.falciparum*, *PLoS Computational Biology*. **3**(2007), e230 (11, 2007).
17. W. F. Doolittle and E. Bapteste, Pattern pluralism and the tree of life hypothesis, *Proceedings of the National Academy of Sciences*. **104**(7), 2043–2049, (2007).
18. C. M. Thomas and K. M. Nielsen, Mechanisms of, and barriers to, horizontal gene transfer between bacteria, *Nat Rev Micro*. **3**(9), 711–721, (2005).
19. C. Pål, B. Papp, and M. J. Lercher, Adaptive evolution of bacterial metabolic networks by horizontal gene transfer, *Nat Genet*. **37**(12), 1372–5 (Dec, 2005).
20. T. Dagan, Y. Artzy-Randrup, and W. Martin, Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution, *Proceedings of the National Academy of Sciences*. **105**(29), 10039–10044, (2008).
21. J. Zhang, Evolution by gene duplication: An update, *Trends Ecol Evol*. **18**(6), 292–

- 298, (2003).
22. M. Nei and A. P. Rooney, Concerted and birth-and-death evolution of multigene families, *Annual Review of Genetics*. **39**(1), 121–152, (2005).
 23. M. Lynch, *The Origins of Genome Architecture*. (Sinauer Associates, Sunderland, MA, 2007).
 24. S. Maslov, K. Sneppen, K. Eriksen, and K. Yan, Upstream plasticity and downstream robustness in evolution of molecular networks, *BMC Evol. Biol.* **4**, 9, (2004).
 25. D. Reichmann, O. Rahat, S. Albeck, R. Megeed, O. Dym, and G. Schreiber, The modular architecture of protein-protein binding interfaces, *Proceedings of the National Academy of Sciences of the United States of America*. **102**(1), 57–62, (2005).
 26. M. Madan Babu, S. A. Teichmann, and L. Aravind, Evolutionary dynamics of prokaryotic transcriptional regulatory networks, *Journal of Molecular Biology*. **358**(2), 614–633, (2006).
 27. E. H. Davidson, *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. (Academic Press, Burlington, USA, 2006).
 28. S. A. Teichmann and M. Babu, Gene regulatory network growth by duplication, *Nature Genetics*. **36**, 492 – 496, (2004).
 29. B. Titz, S. V. Rajagopala, J. Goll, R. Häuser, M. T. McKevitt, T. Palzkill, and P. Uetz, The binary protein interactome of *Treponema pallidum* – the Syphilis spirochete, *PLoS ONE*. **3**(5), e2292, (2008).
 30. J.-C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain, The protein-protein interaction map of *Helicobacter pylori*, *Nature*. **409**, 211–215, (2001).
 31. J. Parrish, J. Yu, G. Liu, J. Hines, J. Chan, B. Mangiola, H. Zhang, S. Pacifico, F. Fotouhi, V. DiRita, T. Ideker, P. Andrews, and R. Finley, A proteome-wide protein interaction map for *Campylobacter jejuni*, *Genome Biology*. **8**(7), R130, (2007).
 32. Y. Shimoda, S. Shinpo, M. Kohara, Y. Nakamura, S. Tabata, and S. Sato, A large scale analysis of protein protein interactions in the nitrogen-fixing bacterium *Mesorhizobium loti*, *DNA Research*. pp. dsm028–, (2008).
 33. G. Butland, J. M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, Interaction network containing conserved and essential protein complexes in *Escherichia coli*, *Nature*. **433**(7025), 531–537, (2005).
 34. S. Sato, Y. Shimoda, A. Muraki, M. Kohara, Y. Nakamura, and S. Tabata, A large-scale protein protein interaction analysis in *Synechocystis* sp. PCC6803, *DNA Research*. **14**(5), 207–216, (2007).
 35. D. J. Lacount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. R. Hesselberth, L. W. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, S. Fields, and R. E. Hughes, A protein interaction network of the malaria parasite *Plasmodium falciparum*, *Nature*. **438**(7064), 103–107 (November, 2005).
 36. S. e. a. Li, A map of the interactome network of the metazoan *c. elegans*, *Science*. **303**, 540–543, (2004).
 37. N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, L. D. Hurst, and M. Tyers, Stratus not altocumulus: A new view of the yeast protein interaction network, *PLoS Biology*. **4**(10), e317 EP –, (2006).
 38. E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J. Girault, B. Goud, J. de Gunzburg, L. Johannes, M. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulou, F. Perez, A. Plessis, C. Rosse, S. Saule, D. Stoppa-Lyonnet,

- A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, and L. Daviet, Protein interaction mapping: a *Drosophila* case study., *Genome Res.* **15**, 376–384, (2005).
39. D. Auerbach, M. Fetchko, and I. Stagljar, Proteomic approaches for generating comprehensive protein interaction maps, *TARGETS*. **2**(3), 85–92, (2003).
 40. J. S. Bader, A. Chaudhuri, J. Rothberg, and J. Chant, Gaining confidence in high-throughput protein interaction networks, *Nat. Biotechnol.* **22**, 78–85, (2004).
 41. J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal, Effect of sampling on topology predictions of protein-protein interaction networks, *Nat. Biotechnol.* **23**, 839–844, (2005).
 42. L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell, Protein-protein interaction networks and biology - what's the connection?, *Nat Biotechnol.* **26**(1), 69–72, (2008).
 43. M. Stumpf, W. Kelly, T. Thorne, and C. Wiuf, Evolution at the system level: the natural history of protein interaction networks, *Trends Ecol Evol.* **22**, 366–373, (2007).
 44. A. Presser, M. B. Elowitz, M. Kellis, and R. Kishony, The evolutionary dynamics of the *Saccharomyces cerevisiae* protein interaction network after duplication, *Proceedings of the National Academy of Sciences*. **105**(3), 950–954, (2008).
 45. B. Bollobás, *Random Graphs*. (Cambridge University Press, 2001), second edition.
 46. A. Barabási and R. Albert, Emergence of scaling in random networks., *Science*. **286**, 509–512, (1999).
 47. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: Simple building blocks of complex networks, *Science*. **298**(5594), 824–827, (2002).
 48. S. Robin, S. Schbath, and V. Vandewalle, Statistical tests to compare motif count exceptionalities, *BMC Bioinformatics*. **8**(1), 84, (2007).
 49. J. J. Daudin, F. Picard, and S. Robin, A mixture model for random graphs, *Statistics and Computing*. **18**(2), 173–183, (2008).
 50. T. Thorne and M. Stumpf, Generating confidence intervals on biological networks, *BMC Bioinformatics*. **8**(1), 467, (2007).
 51. S. Dorogovtsev and J. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*. (Oxford University Press, 2003).
 52. R. Durrett, *Random Graph Dynamics*. Number 20 in Cambridge Series in Statistical and Probabilistic Mathematics, (Cambridge University Press, 2006).
 53. O. Hagberg and C. Wiuf, Convergence properties of some network models., *Bull Math Biol.* **68**, 1275–1291, (2006).
 54. M. Knudsen and C. Wiuf, A Markov chain approach to randomly grown graphs, *Journal of Applied Mathematics*. p. 190836, (2008).
 55. R. M. May, Uses and abuses of mathematics in biology, *Science*. **303**(5659), 790–793, (2004).
 56. G. E. P. Box, Science and statistics, *Journal of the American Statistical Association*. **71**(356), 791–799, (1976).
 57. C. Wiuf, M. Brameier, O. Hagberg, and M. Stumpf, A likelihood approach to analysis of network data, *PNAS*. **103**(20), 7566–7570, (2006).
 58. M. Stumpf, C. Wiuf, and R. May, Subnets of scale-free networks are not scale-free: Sampling properties of networks., *Proc Natl Acad Sci.* **102**, 4221–4224, (2005).
 59. M. Stumpf, P. Ingram, I. Nouvel, and C. Wiuf, Statistical model selection methods applied to biological networks, *Trans. Comp. Sys. Biol.* **3**, 65–77, (2005).
 60. C. Wiuf and M. Stumpf, Binomial subsampling, *Proc Roy Soc A.* **462**, 1181–1195, (2006).
 61. M. P. H. Stumpf and T. Thorne, Multimodel inference of network properties from incomplete data, *J Integr Bioinformatics*. **3**(32), (2007).

- 62. P. Marjoram and S. Tavaré, Modern computational approaches for analysing molecular genetic variation data, *Nat Rev Genet.* **7**(10), 759–770, (2006).
- 63. J. S. Liu, *Monte Carlo Strategies in Scientific Computing.* (Springer-Verlag, New York, 2001).
- 64. E. de Silva, T. Thorne, P. Ingram, I. Agraifoti, J. Swire, C. Wiuf, and M. Stumpf, The effects of incomplete protein interaction data on structural and evolutionary inferences., *BMC Biology.* **4**, 39, (2006).
- 65. M. P. H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, Estimating the size of the human interactome, *Proceedings of the National Academy of Sciences.* pp. 6959–6946, (2008).
- 66. O. Ratmann, C. Andrieu, T. Hinkley, C. Wiuf, and S. Richardson, Model criticism with likelihood-free inference, with an example from evolutionary systems biology, *Proceedings of the National Academy of Sciences.* **to appear**, (2009).

This page intentionally left blank

Chapter 3

Motifs in Biological Networks

FALK SCHREIBER AND HENNING SCHWÖBBERMEYER

*Leibniz Institute of Plant Genetics and Crop Plant Research, Germany
schreibe@ipk-gatersleben.de, schwoebb@ipk-gatersleben.de*

The unprecedented growth in molecular data allows the reconstruction of the structure and dynamics of complex biological processes and systems. To fully understand the function and regulation of complex biological systems it is important to move from the molecular level to the systems level and seek mathematical and computational techniques that can unravel the complexity of the data. Here we characterize the fundamental network building blocks of complex biological systems, and methods that identify and quantify them.

3.1. Introduction

Motifs of statistical significance frequently overlap and form motif complexes. It is unclear if these motif matches represent the basic building blocks of networks and how they differ from functional motifs. To deal with overlapping motifs, the concept of motif themes has been proposed to describe this phenomena.¹

The commonly analysed biological networks represent a static view of all possible interactions. Perhaps the active configurations of the cells have to be analysed to identify the motifs which are really active at a certain point in time from those that emerge solely as a consequence of the network structure.

Current progress in molecular biology, particularly in genome sequencing and high-throughput technologies, have led to an unprecedented growth in data. The availability of detailed molecular data allows the reconstruction of the structure and dynamics of biological processes and systems. This transition from the molecular level to the systems level is necessary for an understanding of the function and regulation of these complex biological systems.^{2,3} In this regard the application of mathematical and computational techniques for the analysis of biological data on the systems level is of great importance due to the complexity of the systems and the wealth of data. A mathematical branch used in modelling complex biological systems is graph theory. The elements of a system are represented as vertices of a graph and the interaction between them are represented as edges. Graph algorithms can then be used to analyse, simulate and visualise the system. Graphs have been used to represent, for example, metabolic, protein-protein interaction and

gene regulatory networks. In these networks entities such as metabolites, proteins or genes are represented by vertices and relationships between entities such as reactions or protein interactions are represented by edges.

The processes of life are highly regulated. A cell, as the smallest entity of life, has the ability to respond to various signals and can adapt to changing conditions of their environment while keeping their internal environment homeostatic. Different mechanisms are recruited for regulation, either short-term regulation by changing the activity of enzymes or long-term regulation by changing the expression level of genes. An important goal of systems biology is to understand the complex regulatory mechanisms of biological systems in detail. The analysis of design patterns of these network regulatory circuits can be useful for understanding the complete systems. Network motifs, patterns of local interconnections (subgraphs), have been described as such basic building blocks of complex networks.⁴ There are several motifs which have been shown to be functionally relevant in biological networks, see Fig. 3.1. Figure 3.2 shows some occurrences of a network motif within a gene regulatory network of yeast (*S. cerevisiae*).

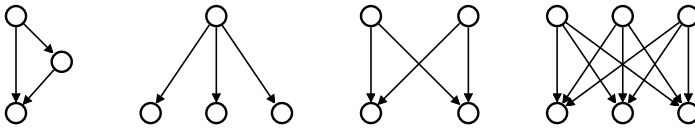


Fig. 3.1. Motifs which have been shown to be functionally relevant in biological networks (from left to right): feed-forward loop motif,^{4–8} single-input motif,^{5,6} bi-fan motif^{4,7,8} and multi-input motif.^{5,7}

3.2. Characterisation of Network Motifs

3.2.1. Definitions

A (directed / undirected) *graph* $G = (V, E)$ consists of a finite set of vertices $V = \{v_1, \dots, v_n\}$ and a finite set of edges $E = \{e_1, \dots, e_m\}$ where each (directed / undirected) edge $e = (v_i, v_j)$ connects two vertices v_i, v_j (in the directed case v_i is the source and v_j is the target). In this chapter we consider directed loop-free (i.e. no edge connects a vertex with itself) graphs. However, the presented method can easily be adapted to other graphs. Let $(e_1, \dots, e_k)$ be a sequence of edges in a graph G . This sequence is called a *walk* if there are vertices $v_0, \dots, v_k$ such that $e_i = (v_{i-1}, v_i)$ for $i = 1, \dots, k$. Two vertices u, v of a graph are *connected* if there exists a walk from vertex u to vertex v . If any pair of different vertices of the graph are connected, the graph is *connected*. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are *isomorphic*, if there exists a bijective mapping between the vertices in V_1 and V_2 , and there is an edge between two vertices of one graph if

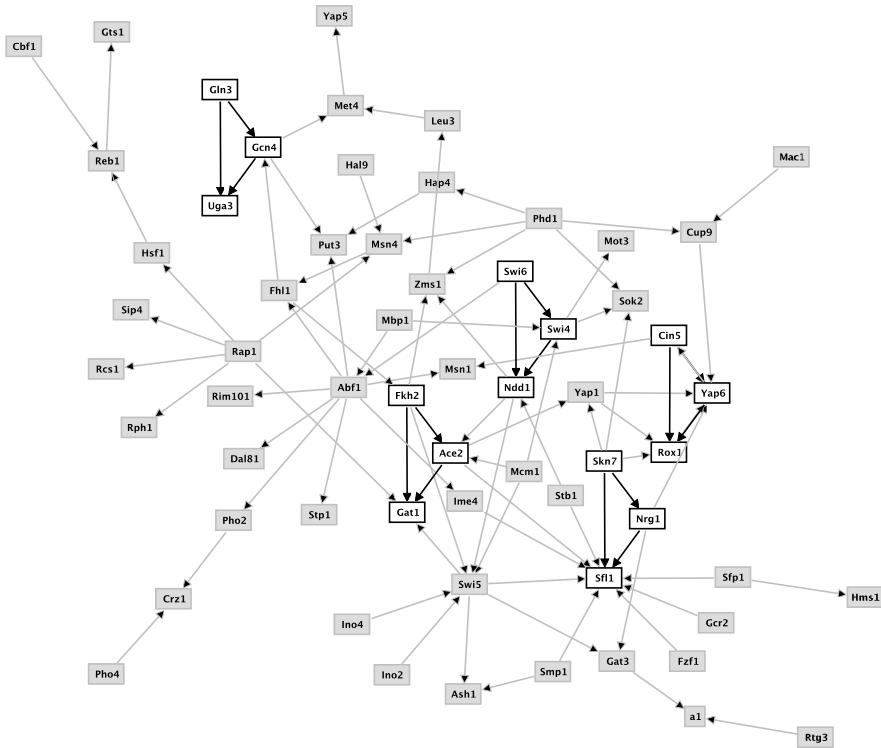


Fig. 3.2. Some occurrences of the feed-forward loop motif (see Fig. 3.1) within a part of the gene regulatory network of yeast (*S. cerevisiae*).

and only if there is an edge between the two corresponding vertices in the other graph. A graph $G' = (V', E')$ is a subgraph of a graph $G = (V, E)$ if $V' \subseteq V$, $E' \subseteq E \cap (V' \times V')$.

A *motif* is a small graph G' . A *match* of a motif within a target graph G is a graph G'' , which is isomorphic to the motif and a subgraph of G , see Fig. 3.3. The frequency of a motif is the number of its matches in the target graph. Different frequency concepts are discussed in Sec. 3.2.4.

3.2.2. Modelling of biological data as graphs

Biological data can often be represented as graphs. To consider two examples, the data from protein-protein interaction experiments can be modelled as a graph with proteins represented by vertices and interactions between proteins modelled as edges. In gene regulatory networks vertices correspond to the DNA sequences (genes) and edges represent interactions between genes (i.e., if the corresponding product of one gene interacts with the promoter of the regulated gene). Figure 3.4

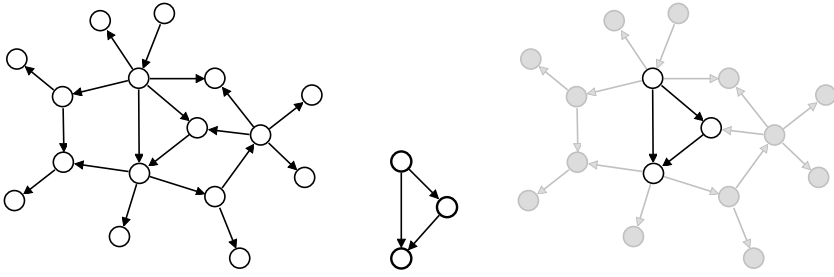


Fig. 3.3. Left: a target graph G . Middle: a motif G' . Right: a match of the motif G' in G .

shows a graph representation of the gene regulatory network in *E. coli*.

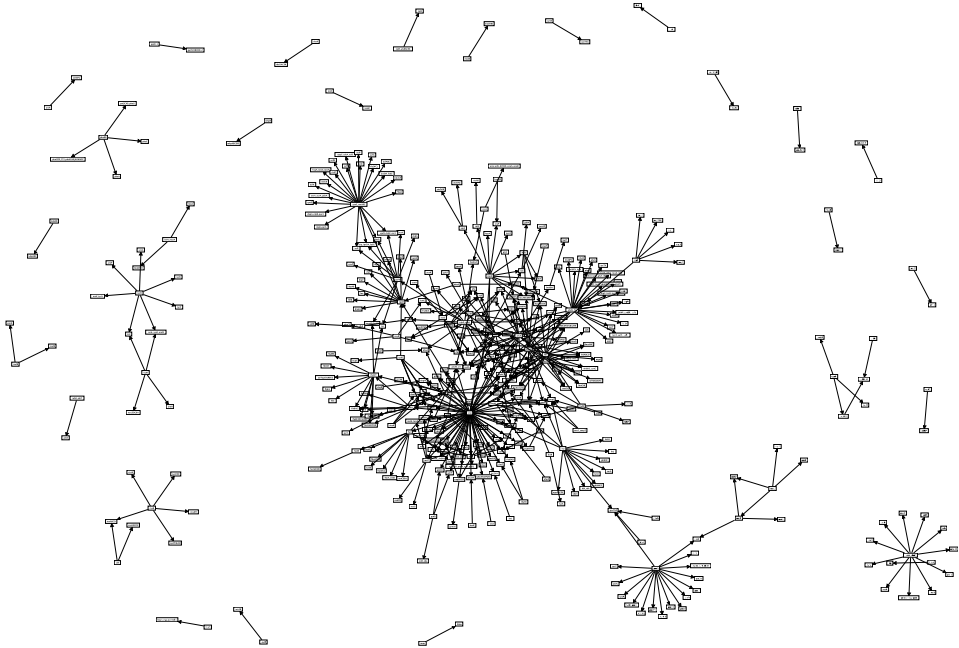


Fig. 3.4. Graph representation of the gene regulatory network in *E. coli*.

3.2.3. Complexity of motif search

Network motif analysis includes several aspects that affect the computational complexity of the task. The number of non-isomorphic graphs grows exponentially with

increasing size, see Table 3.1. Furthermore, there are up to $\binom{|E_m|}{|E_t|}$ matches of a motif $G_m = (V_m, E_m)$ in a graph $G_t = (V_t, E_t)$, where $|E_t|$ represents the number of edges in the target graph and $|E_m|$ is the number of edges in the motif. For the calculation of the statistical significance of network motifs, motif frequencies have to be calculated for a large number of randomised networks.

Despite the high complexity involved in the analysis of network motifs, in practice the search can be executed in reasonable time because typical network motifs are small (three to five vertices) and only a fraction of all possible motifs is supported by a target graph. Furthermore, only some motifs have a high frequency and the majority is less frequent in typical real world networks. Common algorithms and tools for the analysis of network motifs are described in Sec. 3.3.

Table 3.1. Number of non-isomorphic, connected, loop-free undirected and directed graphs for different numbers of vertices.⁹ In case of directed edges, mutual edges (i.e., edges in both directions between two vertices) are allowed.

Vertices	undirected	directed
1	1	1
2	1	2
3	2	13
4	6	199
5	21	9364
6	112	1530843
7	853	880471142
8	11117	1792473955306
9	261080	13026161682466252

3.2.4. Frequency concepts

The frequency of a motif in a particular network is the number of different matches of this motif. There are three reasonable concepts for the determination of the frequency of a motif based on different restrictions on sharing of network elements (vertices or edges) for the matches. These concepts have different properties and are used to analyse different aspects of the motifs, see also Fig. 3.5. Concept $\mathcal{F}_1$ has no restrictions and considers all matches, therefore showing the full potential of a particular motif even if elements of the target graph have to be used several times. Concept $\mathcal{F}_2$ allows the sharing of vertices but not of edges and therefore calculates the number of instances in which a motif has disjoint edges. $\mathcal{F}_2$ shows, for example, in networks where edges represent information flow the number of motif instances that can be ‘active’ at a time. For concept $\mathcal{F}_3$, matches have to be vertex and edge disjoint and can be seen as non-overlapping clusters. This clustering of the target graph allows specific analysis and navigation methods such

as motif-preserving layout of the network.

The restrictions on the reuse of graph elements for concepts $\mathcal{F}_2$ and $\mathcal{F}_3$ have consequences for the determination of motif frequency in the case of overlapping matches, as not all matches can be counted for the frequency. To determine the maximum number of different matches of a motif, the maximum set of non-overlapping matches has to be calculated. This is known as the maximum independent set problem. Since this problem is $\mathcal{NP}$ -complete,¹⁰ usually a heuristic is used to compute a lower bound for the frequency.

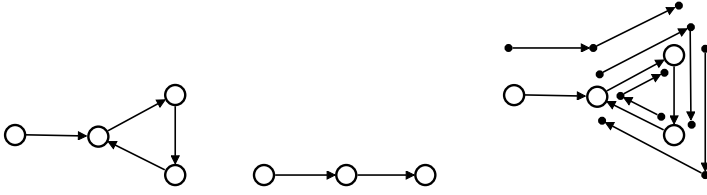


Fig. 3.5. Left: a target graph G . Middle: a motif G' . Right all four matches of the motif G' in G . The application of the different frequency concepts results in a frequency of four for concept $\mathcal{F}_1$, counting all different matches. For $\mathcal{F}_2$ the frequency is two (counting the maximum number of edge-disjoint matches) and for concept $\mathcal{F}_3$ only one match out of the four is valid.

3.2.5. Statistical significance of network motifs

Network motifs are originally defined as patterns of interconnections occurring in networks at numbers that are significantly higher than those in randomised networks⁴ and even though a number of different aspects have been considered,^{5,6,11,12} the statistical significance is still an important property. To calculate the statistical significance of the distribution of motifs in a target network, this distribution is tested against a *random null hypothesis*. For network motifs, the null hypothesis is represented by the distribution of motifs in an ensemble of appropriately randomised networks. Such randomised networks are considered as null hypothesis as their structure is generated by a process free of any type of selection acting on the network's constituent motifs. Rejection of the null hypothesis is taken to represent evidence of functional constraints and design principles that have shaped network architecture at the level of the motifs through selection.^{4,13}

3.2.6. Randomisation algorithm for generation of null model networks

In network motif analysis, a commonly used randomisation algorithm for networks randomly rewires the connections of the network locally.^{14,15} The algorithm reconnects two edges (v_1, v_2) and (v_3, v_4) in such a way that v_1 becomes connected to

v_4 and v_3 to v_2 , provided that none of the newly created edges already exist in the network. This rewiring step is repeated a great number of times to generate a properly randomised network. The essential feature of this algorithm is the preservation of the degree of each vertex. The degree distribution of a network is a characteristic network property and has been used to characterise the large-scale topological structure of biological networks.¹⁶ The applied randomisation algorithm changes the network topology at the local level and preserves the degree distribution at the global level. Therefore, it is believed that this algorithm provides an appropriate null model to calculate the statistical significance of motifs.¹⁵

However, the appropriateness of the randomisation algorithm to represent a random null model has been questioned.¹³ In this paper the authors provide an example where the same motifs have been found in a network created through the process of evolution and a network constructed randomly using a network model which produces a ‘similar’ structure. The statistical relevance of a motif depends on the null model to test for statistical significance. A reformulation of the test for motif significance is required to discriminate functional constraints and design principles from other origins resulting from the network’s construction mechanisms, e.g. spatial clustering.¹³

3.2.7. Calculation of the P-value and Z-score

Statistical significance of motifs for a particular network can be measured by calculating the Z-score and P-value using frequency concept $\mathcal{F}_1$. The Z-score is defined as the difference of the frequency $\mathcal{F}_1$ of this motif in the target network and its mean frequency $\overline{\mathcal{F}_{1,r}}$ in a sufficiently large set of randomised networks, divided by the standard deviation σ_r of the frequency values for the randomised networks,^{4,15} see Eqn. (3.1). The P-value represents the probability P of a motif appearing in a randomised network an equal or greater number of times than in the target network. For a reasonable calculation of the P-value at least 1000 randomised networks have to be considered.¹⁷ Motifs with a P-value less than 0.01 are regarded as statistical significant.⁴ If the number of randomised networks is less than 1000, the P-value is ignored and motifs are considered statistically overrepresented if the Z-score is greater than 2.0.¹⁷

$$\text{Z-score}(m) = \frac{\mathcal{F}_1(m) - \overline{\mathcal{F}_{1,r}(m)}}{\sigma_r(m)}. \quad (3.1)$$

3.3. Methods and Tools for the Analysis of Network Motifs

Different methods and tools have been applied for the analysis of network motifs. Important tools are described in the following Secs. 3.3.1–3.3.3. There are further methods used in the search for network motifs which have been developed for specific questions and are usually not described in detail.^{1,8,12,18–20}

An algorithm for the alignment of motifs was developed to identify motifs derived from families of mutually similar but not necessarily identical patterns.²¹ Publicly available are Matlab scripts¹¹ for motif search which can be found at <http://www.indiana.edu/~cortex/motifs.html>.

3.3.1. *Mfinder*

The *Mfinder* is a software tool for network motif detection in directed and undirected networks.¹⁷ It computes the number of occurrences of a motif of restricted size in the target network (concept $\mathcal{F}_1$) and a uniqueness value, which is a lower bound of the frequency under concept $\mathcal{F}_3$. A value for the frequency under concept $\mathcal{F}_2$ is not calculated. Furthermore, the statistical significance is determined on the basis of the number of occurrences of the motif in randomised networks. The applied randomisation method preserves the degrees of each vertex. The results are presented in a text file and the structure of discovered motifs can be looked up in a motif dictionary.

3.3.2. *Pajek*

Pajek is a program for the analysis and visualisation of large networks.²² It offers the possibility of calculating the frequencies of certain subgraphs like triads and particular tetrads, which are subgraphs with three and four vertices, respectively. Triads can be connected and unconnected and their analysis originates from social network analysis. *Pajek* calculates the number of triads of a network and reports values for the expected frequencies.

3.3.3. *MAVisto*

MAVisto is a tool for the exploration of motifs in biological networks combining a flexible motif search algorithm and different views for the analysis and visualisation of network motifs.²³ It is written in Java and based on *Gravisto*,²⁴ an editor for graphs and a toolkit for implementing graph algorithms. *MAVisto* supports the *Pajek*.net-²² and the GML-format²⁵ and offers graph editor functionality for network manipulation and creation. Furthermore, an advanced force-directed layout algorithm²⁶ is included to generate readable drawings of the network automatically while preserving the layout of motifs where possible. *MAVisto*'s motif search algorithm discovers all motifs of a particular size, which is either given by the number of vertices or by the number of edges. All motifs of this size are analysed and the frequencies for the three different frequency concepts as well as P-value and the Z-score are computed. The measures of statistical significance are obtained by the comparison of motif frequency to randomised versions of the target network. The algorithm for the search is described in detail in Ref. 27. Several views are presented by *MAVisto* in a single interface that assist in the analysis of network motifs:

- (1) The motif table lists information such as the unique network motif label, the size of the motif, some structural properties and the different frequencies together with information about the statistical significance given by the P-value and the Z-score. It allows sorting by all criteria and selecting of motifs to be displayed in the motif view.
- (2) The motif view provides a visual representation of the structure of motifs. Furthermore, it is used to control the display of motif matches in the motif matches view.
- (3) The motif fingerprint represents the motif frequency spectrum of the target network as a diagram. It allows the selection of a column to display the corresponding motif in the motif view.
- (4) The motif matches view provides visual exploration of the occurrences of a motif within the analysed network and supports highlighting of the matches, respectively the covering of network elements by the matches, depending on the applied frequency concept.

The views (1)–(3) allow selection of a motif and the active motif of other perspectives is updated accordingly. This coordination of different views and the possibility of a visual investigation of motif occurrences in networks significantly enhances the explorative power of network motif analysis. In Fig. 3.6 a screenshot of MAVisto is presented showing a step in the analysis of a gene regulatory network.

3.4. Analyses of Motifs in Networks

3.4.1. Analysis of gene regulatory networks

Network motifs have been studied in the well-characterised regulation network of transcriptional interactions in *E. coli*.⁶ In gene regulatory networks, vertices correspond to the DNA sequences (genes) and edges represent interactions between genes (i.e., if the corresponding product of one gene interacts with the promoter of the regulated gene). Three different types of motifs have been identified, the feed-forward loop, the single-input motif and dense overlapping regulons (these are less stringently defined types of multi-input motifs where it is not demanded that every vertex of the output-layer is connected to every vertex of the input layer). Each of the motifs have a specific function in determining gene expression, such as generating temporal expression programs and governing the responses to fluctuating external signals. The whole gene regulatory network can be condensed by merging the nodes of motif instances and representing it by the particular motif. It is proposed that this leads to the identification of the computational layer of the network formed by certain network motifs.⁶

In another study⁵ a gene regulatory network in the eukaryote yeast (*S. cerevisiae*) has been constructed for analysis of its network architecture. Six different types of network motifs with interesting properties have been identified, partially

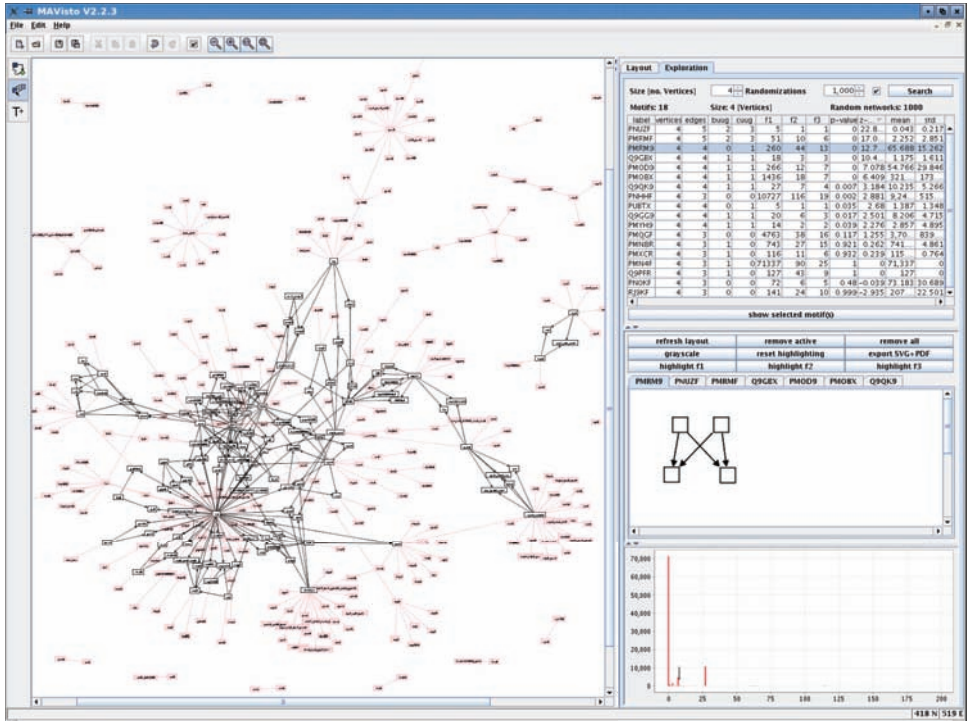


Fig. 3.6. Screenshot of MAVisto showing a step of the analysis of the *E. coli* gene regulatory network. On the left side the analysed network is displayed, on the right side the motif table, the motif view and the motif fingerprint are shown (top to bottom). In the network, elements covered by matches of the motif selected in the motif view are highlighted (black), showing the motif theme of the b-fan motif.

describing sets of related networks. It has been shown that motifs can be used to assemble the gene regulatory network structure of the cell cycle (the sequence of events in a eukaryotic cell that lead from one cell division to the next, divided into four main stages). Furthermore, gene regulators are involved in several processes forming a complex interaction network. For the regulation of the analysed cell cycle, different combinations of regulators are reused at different stages, allowing for a smooth transition to another state. The different substructures of the gene regulatory network are highly interconnected. It is believed that there are higher order transcriptional levels of control within the network, i.e. a hierarchy in the gene regulatory network.⁵

Aside from gene regulatory networks, combinations with other biological networks are also of interest for the analysis of network motifs since these processes do not occur in isolation and are highly interconnected. An integrated network of yeast (*S. cerevisiae*) comprising of gene regulation and protein-protein interactions, modelled by two different types of edges, has been investigated for motifs.²⁸ Besides

the detection of three vertex motifs exhibiting coregulation and complex formation, it was discovered that almost all of the four vertex motifs were combinations of smaller motifs.

3.4.2. *Motifs in cortical networks*

In an analysis of global and local network properties of macaque and cat cerebral cortical networks, significance profiles for three vertex motifs have been further investigated.²⁹ The significance profiles of the two directed networks were highly correlated and were robust against addition, deletion or random switching of connections, suggesting constraints on neocortical development and evolution. The applied randomisation method preserved the degrees of the vertices and the number of two vertex motifs. The comparison to two less stringent methods that preserved (1) only the number of vertices and edges and (2) additionally the degrees of the vertices showed clear differences for some motifs and a low correlation to the stringent significance profile for both networks. However, the significance profiles of the two cortical networks of the macaque and the cat are highly correlated for each of the randomisation method. This indicates that the choice of the network randomisation method is very important in evaluating the local design principles of complex networks.

In another approach,¹¹ network motifs, distinguished between structural and functional motifs, have been investigated in brain networks to study the rules governing their structure. Matches of structural motifs comprise all edges that are present in the network, i.e., they are induced subgraphs (anatomical building blocks), whereas functional motifs are all different motifs that are supported by structural motifs (elementary processing modes of a network). The number of functional motifs of the brain networks is very high compared to random networks, while structural motif number is comparably low. These results are consistent with the hypothesis that highly evolved neural architectures are organised to maximise functional repertoires and to support highly efficient integration of information. The functional motif number has been used as a cost function in an optimisation algorithm to obtain network topologies that resemble real brain networks across a broad spectrum of structural measures. Furthermore, a small set of structural motifs occurring in significantly increased numbers were identified that form a chain of reciprocally connected units. The finding is of interest since this motif type combines two major principles of cortical functional organisation, integration and segregation.

3.4.3. *Analysis of other networks*

The concept of network motifs has been generalised to any type of graph.⁴ Analysis of networks from biochemistry, neurobiology, ecology, and engineering resulted in each case with a distinct set of significant motifs, although some motifs were

shared between different networks. Similar motifs were found in gene regulatory and neuronal networks which both perform biological information processing. It is hypothesised that the motifs occur because of the functional constraints under which the networks have evolved and that motifs can be used for the classification of different network classes.⁴

In a study of networks representing the connection of software class diagrams, the frequency of network motifs has been reasoned to be a consequence of the process of network evolution, thus suggesting a somewhat less relevant role of functionality.³⁰ The analysis of random networks showed that the distribution of motifs depends on the type of network generation mechanism.³¹ Whereas in Erdős-Rényi random networks the frequency is determined by the density of edges, it depends in scale-free networks on the exact topology of the motif.

It is still disputed whether the origin of network motifs in real-world networks is based on spatial properties or whether they arise due to additional functional constraints. For a better understanding of the origin of motifs they have been studied in artificial geometric networks.³² Geometric networks are constructed by placing vertices on a lattice and connecting them with a probability decaying with their distance. This generation process resembles the decay of interactions with increasing distance between vertices in real-world networks. Several invariant measures were found, such as the ratio of feedback and feed-forward loops, which do not depend on network size, dimension, or connectivity function. Furthermore, it was discovered that network motifs in many real-world networks, including social networks and neuronal networks, were not captured solely by these geometric models, supporting the hypothesis that biological network motifs were selected as basic circuit elements with defined information-processing functions.³²

Network motifs have been used as building blocks (coarse-graining units) to generate coarse-grained versions of networks.³³ This approach showed that both biological and electronic networks are *self-dissimilar* and have different network motifs at each level.

3.4.4. Superstructures formed by overlapping motif matches

The gene regulatory network of *E. coli* has been used to study the distribution of motif matches of the feed-forward loop motif and of the bi-fan motif.⁸ For each motif the majority of matches overlap and aggregate into homologous motif clusters. Many of these motif clusters largely overlap with modules of known biological functions within the gene regulatory network. The clusters of overlapping matches of these two motifs aggregate into a superstructure that presents the core or backbone of the network and is assumed to play a central role in defining the global topological organisation. This analysis has introduced distinct topological hierarchies within the *E. coli* transcriptional regulatory network.⁸

The distribution of motif matches has also been analysed in an integrated gene network of yeast (*S. cerevisiae*).¹ In this study the network represented biological

interactions of five different types of the genes and their proteins. The authors described overlapping matches as recurring higher-order interconnection patterns and termed them *network themes*. One example is the *feed-forward theme* – a pair of transcription factors, one regulating the other, and both regulating a common set of target genes that are often involved in the same biological process, see Fig. 3.7. Network themes can be tied to specific biological phenomena and may represent more fundamental network design principles. Furthermore, they provide a useful simplification of complex biological relationships.

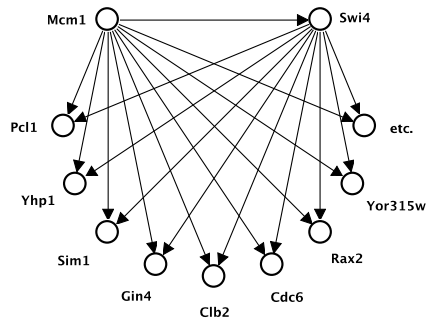


Fig. 3.7. Example of a feed-forward theme of the gene regulatory network of yeast (*S. cerevisiae*) taken from Ref. 1. *Mcm1* regulates *Swi4* and in conjunction they regulate a set of target genes.

The combination of network motifs into larger structures was analysed in a systematic approach that defined *motif generalisations*, families of motifs of different sizes sharing a common architectural theme.³⁴ For the definition of motif generalisations, roles of the vertices were defined according to structural equivalence, e.g. the feed-forward loop motif has three roles: an input node *A*, an output node *C* and an internal node *B* (Fig. 3.8). Motif generalisations are based on the duplication (or multiplication) of one (or more) vertex role(s). Therefore, the feed-forward loop can have three simple generalisations, based on replicating each of the three roles and their connections, as illustrated in Fig. 3.8. It was discovered that networks which share a common motif can have very different generalisations of that motif. Furthermore, the genes of functionally corresponding multi-output feed-forward loop motifs of *E. coli* and yeast (*S. cerevisiae*) gene regulation networks are not evolutionary related, which suggests convergent evolution to the same regulation pattern.³⁴

3.4.5. Dynamic properties of network motifs

The analysis of network motifs has been extended to the investigation of their dynamic properties within biological networks.³⁵ These networks, e.g. gene regulation, signal transduction and neural synapses, are static representations of large-

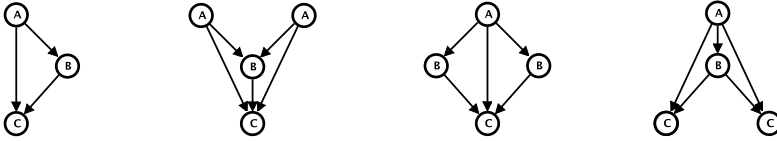


Fig. 3.8. On the left the feed-forward loop motif with labels indicating the roles of the vertices: input (A), internal (B) and output (C). Subsequently, the three simple generalisations of the feed-forward loop motif are shown, replicating the input (A), the internal (B) and the output (C) vertex.

scale dynamic systems with only a particular fraction being active at a time. In this study the dynamic behaviour of three and four vertex network motifs has been systematically determined and related to their distribution in directed networks of gene regulation, developmental regulation, signal transduction and neuronal connections. The dynamic behaviour was characterised by a structural stability score (SSS) that represents the probability of a motif to return to a steady state after small-scale perturbations, defined as intrinsic random fluctuations, or noise, and transient oscillations in activity. Three stability classes have been identified based on the capability of interactions between the vertices of a motif. These classes are stable motifs without feedback interactions ($SSS = 1$), moderately stable motifs with one or two node feedback interactions ($SSS \approx 0.4$) and unstable motifs with feedback interactions between three or more vertices ($SSS < 0.2$). See Fig. 3.9 for examples of motifs of the three classes. The comparison of the frequency of motifs with three and four vertices to random networks of different null models revealed a significant over-representation of motifs with higher structural stability. To exclude impacts of edge numbers on motif frequency from this comparison, the motifs were divided into density groups with equal edge numbers (in software networks it was observed that the most common subgraphs are sparser than less common ones, which are more dense).³⁰ In conclusion, this study proposed that robust dynamical stability of network motifs contributes to biological network organisation and that there is a deep interplay between network structure and system dynamics.³⁵ In a comment on this study it was noted that basic function can be achieved with simple circuits, but if function requires it, complex circuits have evolved along with fine-tuned control mechanisms.³⁶

In another study dealing with dynamic properties of networks, the distribution of feedback and feed-forward loop motifs during information propagation was studied in a signal transduction network.³⁷ The network was constructed based on the signalling pathways and cellular machines in the mammalian hippocampal CA1 neuron. It represents the information flow on the basis of chemical reactions from the response to extracellular ligands to the regulation of components responsible for cellular phenotypic functions. The so-called pseudodynamics of the network

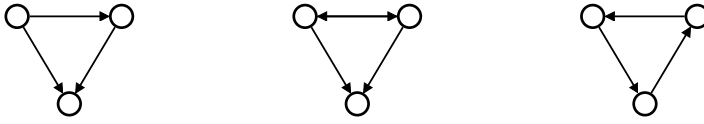


Fig. 3.9. Examples of motifs from the three classes of structural stability. On the left the feed-forward loop represents a structural stable motif as there is no feedback interaction. In the middle a moderately stable motif is shown comprising one mutual edge. On the right a feedback loop is shown as an example of an unstable motif.

(pseudo because it represents propagation of reactions in chemical space rather than time series) was investigated by analysing a series of subnetworks representing the propagation of the signals. At early steps negative feedback loop motifs are abundant or equal to positive feedback loop motifs (see Fig. 3.10), suggesting a barrier to that weak or short-living signals. As the signal propagates, an abundance of positive over negative feedback loop motifs was observed, maybe indicating that signals should persist and be able to evoke a biological response. Furthermore, a higher density of regulatory motifs was found in the middle of the pathways from ligands to cellular machines, indicating that a major portion of the information processing occurs at the ‘centre’ of the network. This study suggests that regulatory motifs are involved in determining cellular choices between homeostasis and plasticity. Cellular systems can be seen as ensembles of different active network configurations and combinations of ligands are likely to produce many more patterns of connectivity, providing a closer view into cellular control mechanisms.



Fig. 3.10. On the left a positive feedback loop with three vertices, on the right a negative feedback loop with four vertices.

3.4.6. Comparison of networks using motif distributions

The protein interaction network of *D. melanogaster* has been classified to a network growth model using the frequencies of particular motifs.³⁸ The model has been selected out of a set of seven network growth models that resemble different mechanisms of network evolution. For this purpose techniques adapted from ma-

chine learning were applied which used the frequencies of motifs as classifiers for the models. Although the network models have similar global network properties, the generated topologies could be distinguished on the basis of the frequency of motifs. In a direct response to this work, difficulties associated with the identification of evolutionary mechanisms that shaped complex networks have been noted.³⁹ Networks underlie varying pressures within their history and the adaptation to these conditions led to changes of the structure. For this reason, the selected network growth model for the *D. melanogaster* protein network captures small-scale features represented by the distribution of network motifs, but some large-scale features are not recapitulated. Moreover, important motifs could be missed by concentrating on motifs where the search is computationally tractable. Available protein interaction networks are not completely correct and they represent a static view of all possible interactions without dynamic information. Nevertheless, it is assumed that the interpretation of a multitude of static data could give clues to dynamic interactions.³⁹

In a similar approach to the classification of the protein interaction network of *D. melanogaster* to a network growth model,³⁸ motifs have been used to select the best fitting model that represents protein interaction networks of *S. cerevisiae* and *D. melanogaster*.⁴⁰ In this work a distance measure for networks has been introduced on the basis of the relative frequency f of subgraphs of size three to five. The distance of two networks was determined by summing up the differences of f for all subgraphs. The model selected by application of this network distance measure showed accordance with the majority of the considered statistical properties for global network structure.

In another approach a method for the classification of complex networks (independent of network size) based on similarities in the local structure has been studied.⁴¹ The classification of directed networks has been based on the statistical significance of motifs; for undirected networks the frequency of motifs relative to random networks was used without considering the statistical significance. For directed networks the Z-scores of motifs with three vertices were used to calculate significance profiles. For undirected networks, the abundance (frequency) of subgraphs with four vertices relative to random networks was used to calculate a subgraph ratio profile. The correlation between significance profiles and ratio profiles was used to cluster the networks into distinct superfamilies. Several of these superfamilies contained networks of different fields with vastly different sizes, e.g. one family contained a network of signal-transduction interactions, a developmental transcription network and a neuronal network. It is currently not verified whether similarity in the profiles is accidental or if the networks have similar key circuit elements because they evolved to perform similar tasks.

The results depend on the suitability of the null hypothesis used to generate the randomised networks for calculation of the statistical significance profile and subgraph ratio profile.¹³ As described in Sec. 3.2.6, the same over-represented motifs were found in real networks and networks generated using a particular network

model. However, by looking at the full subgraph significance profiles there are some motifs which are equally over/under-represented in both the real and the random networks, but some subgraphs show clear differences and allow to distinguish between models and real networks.⁴² Nevertheless, it was proposed that the resolution to distinguish between networks could be increased by the use of higher order subgraphs and a more elaborate null hypothesis could be used to highlight interesting motifs. This increased resolution of higher order subgraphs was confirmed by a comparison of four-vertex motif significance profiles, which put into question the assignment on the basis of three vertex significant profiles, of three networks of developmental regulation, signal transduction and neuronal connections to one superfamily.³⁵

3.4.7. *On the function of network motifs in biological networks*

An analysis of the phylogenetic profiles of genes of different organisms belonging to the class of hemiascomycetes spanning a broad evolutionary range showed that the genes are not subject to any particular evolutionary pressure to preserve the corresponding interaction patterns.¹⁸ There it was discovered that regulatory processes depend on post-transcriptional regulatory mechanisms, rather than on the gene regulation by network motifs. All the examples studied in this analysis highlight the high level of integration of different regulatory mechanisms acting together. Accounting for the various layers of organisation of biological networks seems crucial to correctly identify the functional elements responsible for the information processing.¹⁸

The great majority of motif occurrences are embedded in larger structures and entangled with the rest of the network. This is not taken into account when motifs are considered as isolated functional units. This fact is also not considered by the randomisation process used to generate the null model networks for computing the statistical significance of motifs. Perhaps motifs are a direct consequence of the representation of interaction data in the form of a network.^{18,30} However, the feed-forward loop motif has been shown theoretically and experimentally to have particular kinetic properties that control the temporal program of expression of the target genes.⁴³

The absence of evolutionary pressure for the preservation of particular interaction patterns has also been shown in another study.⁴⁴ This analysis of the evolution of networks revealed that regulatory interactions in motifs are lost and retained at the same rate as the other interactions in the network. There is no bias towards conservation of network motifs by special evolutionary constraints on the constituent elements.

The commonly analysed biological networks represent a static view of all possible interactions. Perhaps the active configurations of the cells have to be analysed to identify the motifs which are really active at a certain point in time from those that emerge solely as a consequence of the network structure.

References

1. L. V. Zhang, O. D. King, S. L. Wong, D. S. Goldberg, A. H. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, and F. P. Rot, Motifs, themes and thematic maps of an integrated *saccharomyces cerevisiae* interaction network, *Journal of Biology*. **4**(2), Epub, (2005).
2. H. Kitano, Systems Biology: A Brief Overview, *Science*. **295**(5560), 1662–1664, (2002).
3. M. Kanehisa and P. Bork, Bioinformatics in the post-sequence era, *Nature Genetics*. **33**, 305–310, (2003).
4. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: Simple building blocks of complex networks, *Science*. **298**(5594), 824–827, (2002).
5. T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*. **298**(5594), 799–804, (2002).
6. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nature Genetics*. **31**(1), 64–68, (2002).
7. G. C. Conant and A. Wagner, Convergent evolution of gene circuits, *Nature Genetics*. **34**(3), 264–266, (2003).
8. R. Dobrin, Q. K. Beg, A.-L. Barabási, and Z. N. Oltvai, Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network, *BMC Bioinformatics*. **5**(1), 10, (2004).
9. F. Harary and E. M. Palmer, *Graphical Enumeration*. (Academic Press, New York, 1973).
10. M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. (W.H. Freeman and Company, New York, 1979).
11. O. Sporns and R. Kötter, Motifs in brain networks, *PLoS Biology*. **2**(11), e369, (2004).
12. S. Wuchty, Z. N. Oltvai, and A.-L. Barabási, Evolutionary conservation of motif constituents in the yeast protein interaction network, *Nature Genetics*. **35**(2), 176–179, (2003).
13. Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, Comment on “Network motifs: simple building blocks of complex networks” and “Superfamilies of evolved and designed networks”, *Science*. **305**(5687), 1107c, (2004).
14. S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks, *Science*. **296**, 910–913, (2002).
15. S. Maslov, K. Sneppen, and U. Alon. Correlation profiles and motifs in complex networks. In eds. S. Bornholdt and H. G. Schuster, *Handbook of Graphs and Networks: From the Genome to the Internet*, pp. 168–198. Wiley-VCH, (2003).
16. A.-L. Barabási and Z. N. Oltvai, Network biology: understanding the cell’s functional organization, *Nature Reviews Genetics*. **5**(2), 101–113, (2004).
17. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Mfinder tool guide. Technical report, Department of Molecular Cell Biology and Computer Science & Applied Mathematics, Weizman Institute of Science, (2002).
18. A. Mazurie, S. Bottani, and M. Vergassola, An evolutionary and functional assessment of regulatory network motifs., *Genome Biology*. **6**(4), R35, (2005).
19. H. S. Moon, J. Bhak, K. H. Lee, and D. Lee, Architecture of basic building blocks in protein and domain structural interaction networks, *Bioinformatics*. **21**(8), 1479–

- 1486, (2005).
20. M. Reigl, U. Alon, and D. B. Chklovskii, Search for computational modules in the *C. elegans* brain., *BMC Biology*. **2**(1), 25, (2004).
21. J. Berg and M. Lässig, Local graph alignment and motif search in biological networks, *Proc. Natl. Acad. Sci. USA*. **101**(41), 14689–14694, (2004).
22. V. Batagelj and A. Mrvar. Pajek - analysis and visualization of large networks. In eds. M. Jünger and P. Mutzel, *Graph Drawing Software*, pp. 77–103. Springer, (2004).
23. F. Schreiber and H. Schwöbbermeyer, MAVisto: a tool for the exploration of network motifs, *Bioinformatics*. **21**(17), 3572–3574, (2005).
24. C. Bachmaier, F. J. Brandenburg, M. Forster, M. Raitner, and P. Holleis. Gravisto: Graph visualization toolkit. In *Proceedings of the International Symposium on Graph Drawing (GD 2004)*, vol. 3383, *Lecture Notes in Computer Science*, pp. 502–503. Springer, (2005).
25. M. Himsolt, Graphlet: design and implementation of a graph editor, *Software - Practice and Experience*. **30**(11), 1303–1324, (2000).
26. T. Fruchterman and E. Reingold, Graph drawing by force-directed placement, *Software - Practice and Experience*. **21**(11), 1129–1164, (1991).
27. F. Schreiber and H. Schwöbbermeyer, Frequency concepts and pattern detection for the analysis of motifs in networks, *Transactions on Computational Systems Biology*. **3**, 89–104, (2005).
28. E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit, Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction, *Proc. Natl. Acad. Sci. USA*. **101**(16), 5934–5939, (2004).
29. S. Sakata, Y. Komatsu, and T. Yamamori, Local design principles of mammalian cortical networks, *Neuroscience Research*. **51**(3), 309–315, (2005).
30. S. Valverde and R. V. Sole, Network motifs in computational graphs: A case study in software architecture, *Physical Review E*. **72**(2):026107, (2005).
31. S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, Subgraphs in random networks, *Physical Review E*. **68**(2):026127, (2003).
32. S. Itzkovitz and U. Alon, Subgraphs and network motifs in geometric networks, *Physical Review E*. **71**(2):026117, (2005).
33. S. Itzkovitz, R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz, and U. Alon, Coarse-graining and self-dissimilarity of complex networks, *Physical Review E*. **71**(1):016127, (2005).
34. N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, Topological generalizations of network motifs, *Physical Review E*. **70**(3):031909, (2004).
35. R. J. Prill, P. Iglesias, and A. A. Levchenko, Dynamic properties of network motifs contribute to biological network organization, *PLoS Biology*. **3**(11), e343, (2005).
36. J. Doyle and M. Csete, Motifs, control, and stability, *PLoS Biology*. **3**(11), e392, (2005).
37. A. Ma’ayan, S. L. Jenkins, S. Neves, A. Hasseldine, E. Grace, B. Dubin-Thaler, N. J. Eungdamrong, G. Weng, P. T. Ram, J. J. Rice, A. Kershenbaum, G. A. Stolovitzky, R. D. Blitzer, and R. Iyengar, Formation of Regulatory Patterns During Signal Propagation in a Mammalian Cellular Network, *Science*. **309**(5737), 1078–1083, (2005).
38. M. Middendorff, E. Ziv, and C. H. Wiggins, Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network, *Proc. Natl. Acad. Sci. USA*. **102**(9), 3192–3197, (2005).
39. J. J. Rice, A. Kershenbaum, and G. Stolovitzky, Lasting impressions: Motifs in protein-protein maps may provide footprints of evolutionary events, *PNAS*. **102**(9),

- 3173–3174, (2005).
40. N. Pržulj, D. G. Corneil, and I. Jurisica, Modeling interactome: scale-free or geometric?, *Bioinformatics*. **20**(18), 3508–3515, (2004).
 41. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, Superfamilies of evolved and designed networks, *Science*. **303**(5663), 1538–1542, (2004).
 42. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon, Response to comment on “Network motifs: Simple building blocks of complex networks” and “Superfamilies of evolved and designed networks”, *Science*. **305**(5687), 1107d, (2004).
 43. S. Mangan, A. Zaslaver, and U. Alon, The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks, *J. Mol. Biol.*. **334**(2), 197–204, (2003).
 44. M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, Structure and evolution of transcriptional regulatory networks, *Curr. Opin. Struct. Biol.* **14**(3), 283–291, (2004).

Chapter 4

Bayesian Analysis of Biological Networks: Clusters, Motifs, Cross-Species Correlations

JOHANNES BERG AND MICHAEL LÄSSIG

*Institut für Theoretische Physik, Universität zu Köln, Germany
berg@thp.uni-koeln.de, lassig@thp.uni-koeln.de*

Detecting functionality in biological networks is a major goal of systems biology. Such networks consist of functional units in an effectively random background, so we need statistical models and algorithms to discriminate both parts. In this chapter, we develop a statistical theory of network topology, using the evolutionary dynamics of nodes and links to distinguish functional from random parts. We discuss three particular cases: clusters within a network, repetitive network motifs and cross-species correlations between networks, with examples from protein interaction networks, transcriptional regulation networks and co-expression networks.

4.1. Introduction

The complexity of an organism is only weakly linked with its number of genes. *Homo sapiens* has about 25,000 genes and the roundworm *C. elegans* about 19,000,^{1,2} despite the different levels of complexity. Not only are the gene numbers similar, the genes themselves are frequently shared across species. Even distantly related organisms have a high fraction of genes which stem from a common ancestor (orthologues): more than 90% of genes are shared between human and mouse and at least 30% of genes of the yeast *S. cerevisiae* have orthologues in human.³

This result is an important outcome of the recent genome sequencing projects. It has put the spotlight on the *interactions* between genes: changes in the complex networks of gene regulation or in the interactions between proteins may be a major cause of phenotypic variation, more so than changes in the genes themselves.⁴ The molecular basis of these interactions includes specific binding sites on regulatory DNA and binding domains in proteins. Binding sites can change quickly, generating new interactions or deleting old ones.^{5–8}

The resulting interest in biological interactions has been matched by the development of novel experimental techniques to measure protein-DNA interactions and protein-protein interactions. In particular, *high-throughput* methods have been developed, facilitating measurements on a genome-wide scale rather than for individual genes. Some of the ingenious methods of experimentally determining biological

interactions will be briefly reviewed in the next section.

This experimental development is akin to the transition from sequencing small parts of the DNA of an organism to the determination of full genomes. The growth of sequencing capabilities has been driving the development of computational methods for sequence analysis for the past three decades. Virtually all methods for sequence analysis rely on statistics as a tool to infer function. Examples are the detection of genes, or of regulatory modules, or the identification of correlations between evolutionarily related sequences.⁹

The corresponding development of computational network biology is still in its infancy. New tools will be required to address specific issues of biological networks. These are characterised by a peculiar interplay of stochasticity and function, and in many ways epitomise our current lack of understanding of biological systems. With this caveat, the point of view we take in this article is that statistics will again play a decisive role in our understanding of network biology. We will also point out some currently available links between network statistics and function. The merit of a statistical approach may not seem obvious from an engineering perspective, where networks are seen as deterministic processing machines producing a well-defined input-output relation. Indeed, biological networks sometimes work in a surprisingly deterministic way: for example, a network of a few dozen major genes generates a well-defined spatiotemporal development pattern in the eukaryotic embryo. However, the underlying network structures are fundamentally stochastic since they arise from the manifold tinkering and feedback processes of biological evolution. Explaining deterministic function from a stochastic evolution requires a statistical, dynamical theory.

One important aspect of this challenge is to predict different functional units in networks. Different functions are reflected in different evolutionary dynamics, and hence in different statistical characteristics of network parts. In this sense, the *global statistics* of a biological network, e.g., its connectivity distribution, provides a background, and *local deviations* from this background signal functional units. Thus, in the computational analysis of biological networks, we typically have to discriminate between different statistical models governing different parts of the dataset. The nature of these models depends on the biological question asked. We illustrate this rationale here with three examples: the identification of functional parts as highly connected *network clusters*, the search for *network motifs*, which occur in similar forms at different places in the network, and the analysis of *cross-species network correlations*, which reflect evolutionary dynamics between species.

4.2. Measuring Biological Networks

A wide range of experimental methods has been developed to measure interactions between proteins, interactions between proteins and regulatory DNA, and expression levels of genes. Only a brief review is possible here.

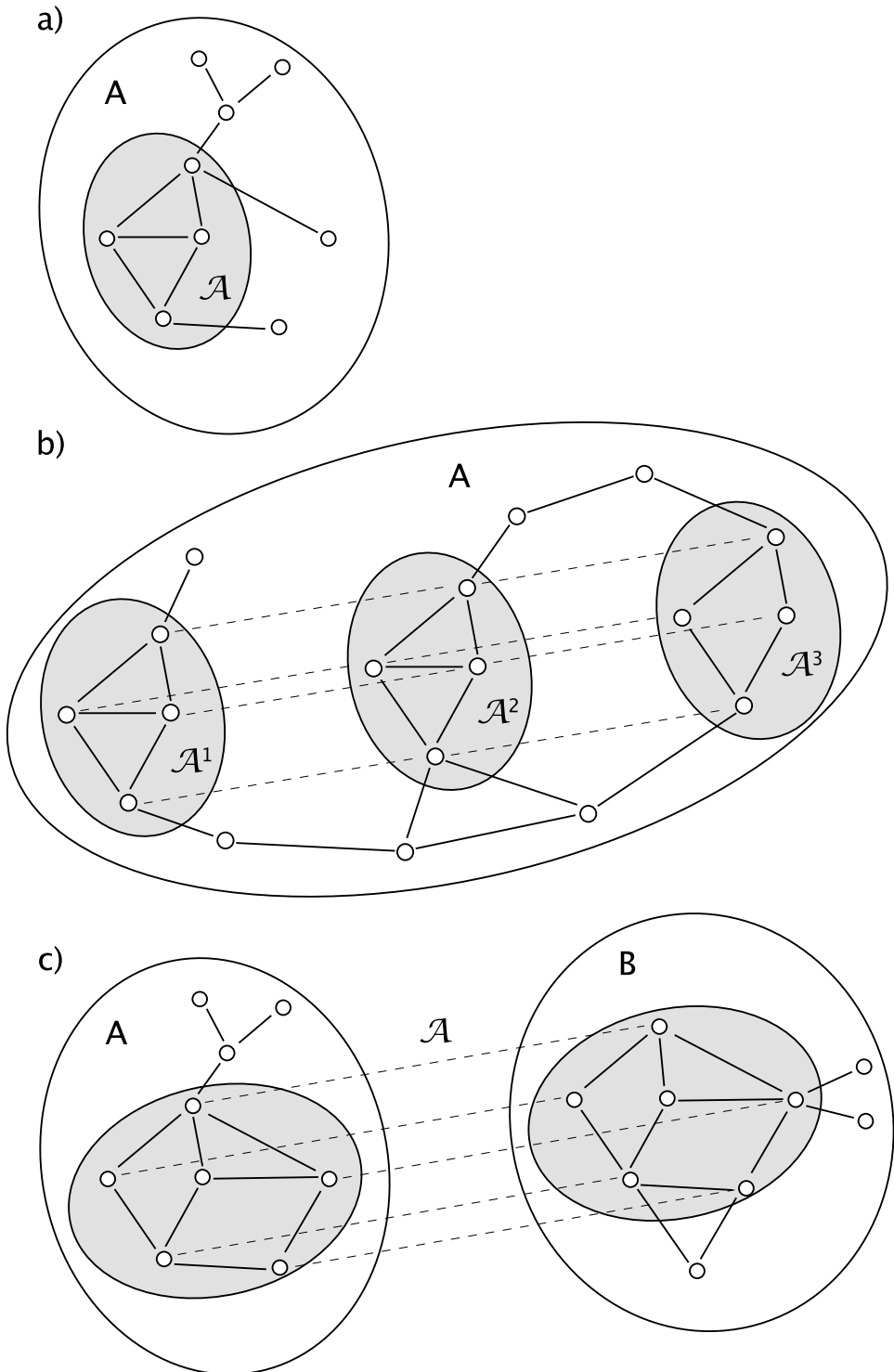


Fig. 4.1. Deviation from a uniform global statistics in biological networks. (A) A network cluster is distinguished by an enhanced number of intra-cluster interactions. (For details see Sec. 4.4.) (B) A network motif is a set of subgraphs with correlated interactions. (See Sec. 4.5.) In a limiting case, all subgraphs have the same topology. (C) Cross-species correlations characterise evolutionarily conserved parts of networks. (See Sec. 4.6.)

In the yeast two-hybrid (Y2H) method, the pairwise interaction between two proteins is tested by creating two fusion proteins.¹⁰ One protein is constructed with a DNA-binding domain attached to its end, and its potential binding partner is fused to an activation domain. If the two proteins interact, the binding will form a transcriptional activator (generally consisting of a DNA-binding domain and an activation domain). The presence of an intact activator leads to the transcription of an easily detectable reporter gene. (The reporter gene may for instance produce a fluorescent protein.) In principle, the amount of the reporter gene produced can serve as a measure of the affinity between the two proteins. The Y2H method has been used to measure the protein interaction networks of yeast,¹⁰ *C. elegans*,¹¹ *D. melanogaster*¹² and human.¹³ The Y2H datasets are known to contain a large number of false positive and false negative results. False negatives arise when the fusion proteins fail to localise in the yeast nucleus, or fail to fold properly once the new domains are attached. False positives may be linked to high expression levels of the hybrid in yeast, which are never reached *in vivo*.

Alternative approaches include pull-down assays, where one protein type is immobilised on a gel, and ‘pulls down’ binding partners from a solution. Binding partners may then be identified by various tags. Mass spectrometry is also used to identify the interacting protein pairs identified by such an affinity analysis.¹⁴ While more accurate than the Y2H method, these approaches have not yet been scaled up to provide high throughputs.

Binding of proteins, specifically transcription factors, to regulatory DNA has long been investigated by electrophoresis, where the motility of a DNA fragment is altered by a protein bound to it. Chromatin immunoprecipitation (ChIP) is an alternative procedure, which uses specific antibodies to isolate a protein and then amplifies DNA that may have been isolated together (co-precipitated) with the protein. By running many such experiments in parallel on a microarray, this method can be scaled up to high throughputs (ChIP-on-chip¹⁵).

Gene expression levels can be measured on DNA microarrays, densely packed samples of known nucleotides, each a few tens of base pairs long. Currently more than 10^6 of such samples, or probes, can be placed on a single microarray. The array is then washed with a fluorescently labelled sample. Binding of DNA in the sample to complementary DNA on the probe can be detected under a microscope from the resulting fluorescence pattern. Genome-wide expression levels can thus be measured on a single array. Many other applications of microarrays are being developed – for instance microarrays to measure interactions between transcription factors and regulatory DNA. DNA microarrays are also making major inroads as diagnostic tools, from characterising the microbial communities in dentistry¹⁶ to the early detection of cancer.¹⁷

4.3. Random Networks in Biology

Randomly generated networks are very useful for analysing simple characteristics of biological networks. For instance, typical distances on a randomly generated network generally scale logarithmically with the number of network nodes. Finding such short distances in biological network data as well is therefore not a surprising result and does not require a biological explanation. Another frequent observation in biological networks is a distribution of node connectivities with a broad tail, which is shared by specific ensembles of random networks. This has motivated a number of statistical models explaining the connectivity distribution in terms of the underlying evolutionary dynamics.^{18–20} Thus, ensembles of random networks can be tuned to fit certain characteristics of biological network data. Does that mean the actual network *is* random? This is clearly not the case: other observables may differ from what is expected in the random network ensemble, and we will see that these deviations from the ‘null model’ are particularly interesting as signals of biological function. Hence, random network ensembles play an important role in quantifying the most unbiased background statistics of a ‘functionless’ network. Their choice is a subtle issue: it has to be motivated by what we consider to be unimportant for the biological function in question. Let us now turn to a few such models.

A network is specified by its adjacency matrix $\mathbf{a} = (a_{ii'})$. For binary networks $a_{ii'} = 1$ if there is a link between nodes i and i' , and $a_{ii'} = 0$ if there is no link. Networks with undirected links are represented by a symmetric adjacency matrix. The *in* and *out connectivities* of a node, $k_i^+ = \sum_{i'} a_{i'i}$ and $k_i^- = \sum_{i'} a_{ii'}$, are defined as the number of in- and outgoing links, respectively. The total number of directed links is given by $K = \sum_{i,i'} a_{ii'}$.

To focus on a specific part of the network we define an ordered subset $\mathcal{A}$ of n nodes $\{r_1, \dots, r_n\}$ (see Fig. 4.1A). The subset $\mathcal{A}$ induces a *pattern* $\hat{\mathbf{a}}(\mathcal{A})$ on the network, represented by the restricted adjacency matrix containing only links internal to node subset $\mathcal{A}$. $\hat{\mathbf{a}}$ is thus an $n \times n$ matrix with entries $\hat{a}_{ij} = a_{r_i r_j}$ ($i, j = 1, \dots, n$). Together, the subset of nodes $\mathcal{A}$ and its pattern $\hat{\mathbf{a}}(\mathcal{A})$ form a *subgraph*.

The simplest ensemble of random networks is generated by connecting all pairs of nodes independently with the same probability w . Given a subset of nodes $\mathcal{A}$, the probability of generating pattern $\hat{\mathbf{a}}$ is then given by $P_0(\mathbf{a}) = \prod_{i,i' \in \mathcal{A}} (1-w)^{1-a_{ii'}} w^{a_{ii'}}$ (for undirected networks the sum is restricted to $i \leq i'$). This well-known ensemble, named after the pioneers of graph theory P. Erdős and A. Rényi, leads to a Poisson distribution of connectivities. The only free parameter of the Erdős–Rényi (ER) model, the link probability w between a given pair of nodes, can be tuned so that typical graphs taken from the ER ensemble have the same number of links as the empirical data. If the subset of nodes $\mathcal{A}$ contains all $n = N$ nodes of the network, $w = K/N^2$. Considering connected subgraphs with $n < N$, w will in general be higher than K/N^2 . Then the value of w can be determined by generating all

connected subgraphs of size n from the empirical dataset and choosing w such that the average number of links in the ER model equals the average number of links in connected subgraphs in the data.

However, in biological networks the connectivity distribution often differs markedly from that of the Erdős–Rényi model. If we have reasons to assume that a biological function is not tightly linked to connectivity at the level of individual nodes, we should include the connectivity distribution in our null model. Indeed, we can easily construct a random ensemble matching the connectivity distribution of the dataset. In this ensemble, the probability $w_{ii'}$ of finding a link between a pair of nodes i, i' depends on the connectivities of the nodes. Assuming links between different node pairs to be uncorrelated, a given subset of nodes $\mathcal{A}$ has a pattern $\hat{\mathbf{a}}$ with probability

$$P_0(\hat{\mathbf{a}}) = \prod_{i, i' \in \mathcal{A}}^n (1 - w_{ii'})^{1 - a_{ii'}} w_{ii'}^{a_{ii'}}. \quad (4.1)$$

For $n = N$, when $\mathcal{A}$ includes the entire network, the probability of finding a directed link between nodes i and i' is approximately $w_{ii'} = k_{r_i}^- k_{r_{i'}}^+ / K$, that of an undirected link $w_{ii'} = k_{r_i} k_{r_{i'}} / K$.²¹ Furthermore, if we impose the constraint that the null model describe the statistics of a *connected* dataset, the probabilities in Eqn. (4.1) are increased by a factor that can be determined from the data as described above. The null model constructed in this way is maximally unbiased with respect to all patterns in the dataset beyond its connectivity distribution.

4.4. Network Clusters

A first trace of functionality in biological networks is strong inhomogeneities in their link statistics, which are not captured by the null model. Examples are aggregates of several proteins held together by mutual interactions, which show up as highly connected clusters in protein interaction networks, and sets of co-regulated genes (for instance co-regulated by an oncogene),²² leading to clusters in co-expression networks. How can we identify these clusters statistically?

Clusters are subgraphs with a significantly increased number of internal links compared to the background of the network, see Fig. 4.1A. The feature that distinguishes clusters is the number of internal links,

$$L(\hat{\mathbf{a}}) = \sum_{i, i' \in \mathcal{A}}^n \hat{a}_{ii'}. \quad (4.2)$$

The statistics of clusters is then described by an ensemble

$$Q_\sigma(\hat{\mathbf{a}}) = Z_\sigma^{-1} \exp[\sigma L(\hat{\mathbf{a}})] P_0(\hat{\mathbf{a}}) \quad (4.3)$$

of the same form as Eqn. (4.1), but with a bias towards a high number of internal links. The average number of internal links is determined by the value

of the link reward σ . We have introduced the normalisation factor $Z_\sigma = \prod_{ii'}^n \sum_{\hat{a}_{ii'}=0,1} \exp[\sigma L(\hat{\mathbf{a}})] P_0(\hat{\mathbf{a}})$, which ensures that $Q_\sigma(\hat{\mathbf{a}})$ summed over all patterns $\hat{\mathbf{a}}$ gives unity.

Is a given pattern $\hat{\mathbf{a}}$ more likely to be part of a cluster as described by the model (4.3), or is it more likely to be part of the background described by the null model (4.1)? To address this question, we define the so-called *log-likelihood score*

$$S(\mathcal{A}, \sigma) = \log \left(\frac{Q_\sigma(\hat{\mathbf{a}})}{P_0(\hat{\mathbf{a}})} \right) = \sigma L(\hat{\mathbf{a}}(\mathcal{A})) - \log Z_\sigma . \quad (4.4)$$

A positive score results if it is more likely for the pattern $\hat{\mathbf{a}}(\mathcal{A})$ to arise in the model describing clusters than in the alternative null model. High scores indicate strong deviations from the null model. Of course this is an attractive property for the algorithmic search for deviations from the null model. As shown in the appendix, the form of the score is related in a simple way to the probability that pattern $\hat{\mathbf{a}}$ comes from the model describing clusters.

Patterns with a high score are *bona fide* clusters. The first term of the score weighs the total number of links. As expected, a pattern with many internal links yields a high score. The second term acts as a threshold and assigns a negative score to a pattern with too few internal links. This term takes into account the connectivities of the nodes: highly connected nodes have more internal links already in the null model. Node subsets with highly connected nodes tend to give lower scores. The score thus goes beyond simple measures of clustering, such as the number of internal links, and provides a statistical basis for cluster detection. Given the scoring parameter σ , the maximum-score node subset $\mathcal{A}^*(\sigma)$ is defined by

$$\mathcal{A}^*(\sigma) = \operatorname{argmax}_{\mathcal{A}} S(\mathcal{A}, \sigma) . \quad (4.5)$$

At this point, the scoring parameter σ is a free parameter, whose value needs to be inferred from the data. This can be done by applying the principle of maximum likelihood: σ is determined by the requirement that the model describing clusters (4.3) optimally describes the statistics of the maximum-score pattern. For a given pattern $\hat{\mathbf{a}}$, the optimal fit is defined by the so-called *maximum likelihood value* $\sigma^* = \operatorname{argmax}_{\sigma} Q_\sigma(\hat{\mathbf{a}}(\mathcal{A}))$, which maximises the likelihood of generating pattern $\hat{\mathbf{a}}(\mathcal{A})$ under the model (4.3). Since $\log(x)$ is a monotonously increasing function, the maximum likelihood value σ^* coincides with the maximum of the log-likelihood score (4.4) over σ . The maximum-score node subset at the optimal scoring parameter is then determined by the joint maximum of the score over $\mathcal{A}$ and σ

$$S(\mathcal{A}^*, \sigma^*) = \max_{\sigma} S(\mathcal{A}^*(\sigma), \sigma) = \max_{\mathcal{A}, \sigma} S(\mathcal{A}, \sigma) . \quad (4.6)$$

One can easily show that the maximum-likelihood value of σ sets the expected number of links in the ensemble Q_{σ^*} equal to the actual number of links in pattern $\hat{\mathbf{a}}^*$: setting the derivative of Eqn. (4.4) with respect to σ equal to zero gives

$$\langle L(\hat{\mathbf{a}}) \rangle_{Q_{\sigma^*}} = L(\hat{\mathbf{a}}^*) . \quad (4.7)$$

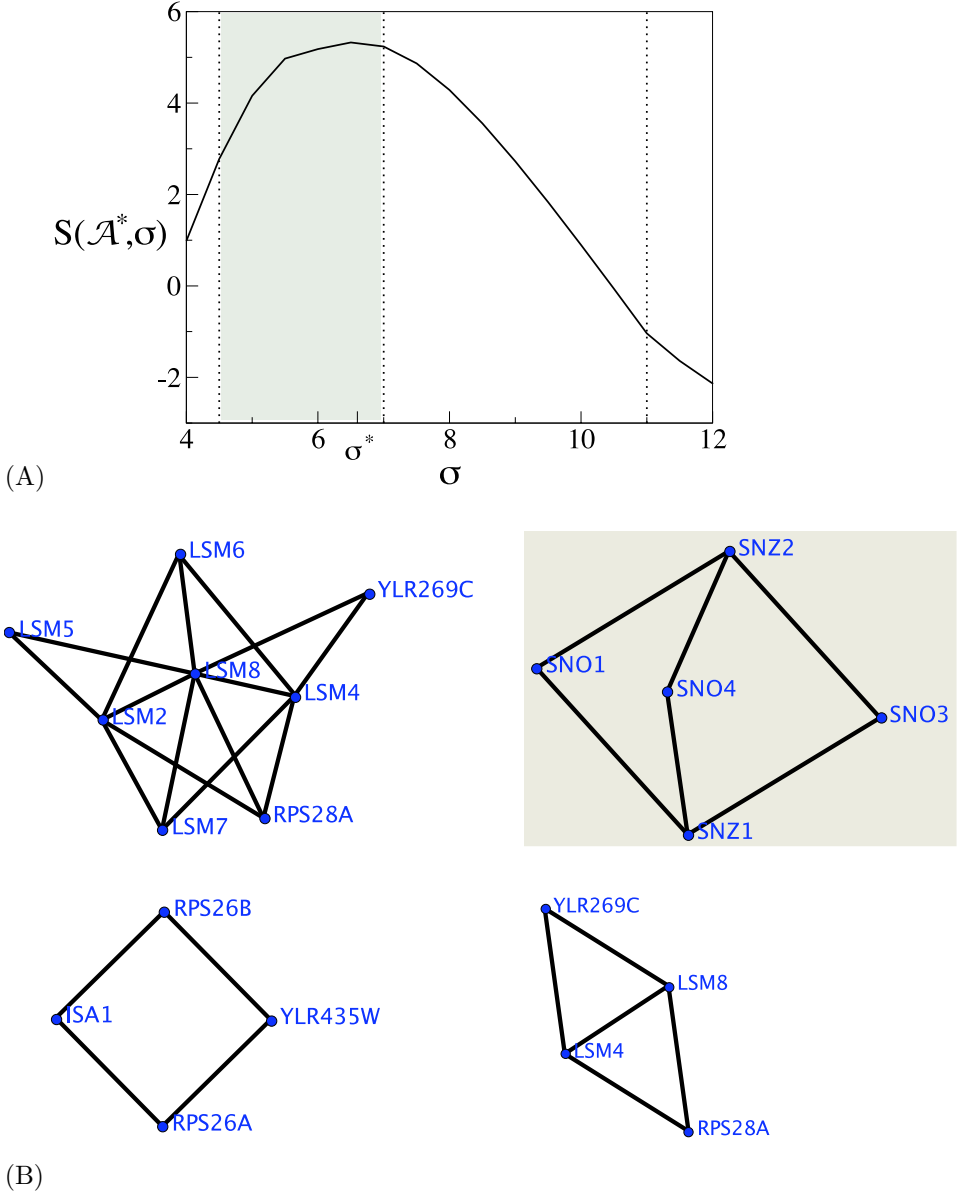


Fig. 4.2. Scoring clusters in protein interaction networks. (A) The score S of the maximum-score node subset $\mathcal{A}^*(\sigma)$ is shown as a function of the scoring parameter σ . The dotted lines indicate the values of σ where the maximum-score node subset changes. The maximum of the score with respect to σ indicates the optimal scoring parameter $\sigma^* = 6.6$. The grey region $4.25 < \sigma < 7$ indicates the values where $\mathcal{A}^*(\sigma) = \mathcal{A}^*(\sigma^*)$. (B) The maximum-score subgraphs for $\sigma < 4.25$, $4.25 < \sigma < 7$, $7 < \sigma < 11$, $\sigma > 11$ (left to right). The (unique) subgraph resulting from the optimal scoring parameter is highlighted in grey. The maximum-score subgraphs for $7 < \sigma < 11$ and for $\sigma > 11$ are distinguished by the connectivities of their nodes, with the latter having a higher average connectivity. This accounts for the former having a higher score for $7 < \sigma < 11$ despite the smaller number of internal links.

4.4.1. Clusters in protein interaction networks

We use the scoring function (4.4) to identify clusters in the protein interaction network of yeast, namely the high-throughput dataset of Uetz *et al.*¹⁰ At a given value of the scoring parameter σ , the maximum-score node subset $\mathcal{A}^*(\sigma)$ is identified using a simple Monte Carlo algorithm. At different values of σ , different node subsets $\mathcal{A}^*(\sigma)$ yield the highest score (compared to all other node subsets). The resulting subgraphs are shown in Fig. 4.2A. At low values of σ , subgraphs with many nodes, but comparatively few internal interactions per node, yield the highest score. At high values of σ , subgraphs with many internal interactions are favoured. However these subgraphs tend to be small. The interplay between subgraph size and internal connectivity leads to a joint score maximum over $\mathcal{A}$ and σ at the optimal scoring parameter $\sigma^* = 6.6$, see Fig. 4.2A.

The maximum-score cluster $\mathcal{A}^* \equiv \mathcal{A}^*(\sigma^*)$ consists of the proteins SNZ1, SNZ2, SNO1, SNO3, and SNO4, highlighted in grey in Fig. 4.2B. The proteins in this cluster have a common function; they are involved in the metabolism of pyridoxine and in the synthesis of thiamin.^{23,24} Furthermore, SNZ1 and SNO1 have been found to be co-regulated and their mRNA levels increase in response to starvation for amino acids A, U, and Trp.²⁵

4.5. Network Motifs

The topology of a subgraph may be associated with a specific function. A possible example is a feed-forward loop acting as a high-frequency filter in a regulatory network.²⁶ If such a function is required repeatedly in different parts of the network, there is selection pressure for the creation and maintenance of similar topologies in different parts of the network. Such *network motifs*^{26,27} are families of subgraphs distinguished from the null model by mutual correlations between subgraphs, see Fig. 4.1B.

To quantify these correlations, we need to specify the parts of the network with correlated patterns. We define a *graph alignment* $\mathcal{A}$ by a set of several node subsets $\mathcal{A}^\alpha$ ($\alpha = 1, \dots, p$), each containing the same number of n nodes, and a specific order of the nodes $\{r_1^\alpha, \dots, r_n^\alpha\}$ in each node subset. An alignment associates each node in a node subset with exactly one node in each of the other node subsets. The alignment can be visualised by n ‘strings’, each connecting p nodes as shown in Fig. 4.1B.

An alignment specifies a pattern $\hat{\mathbf{a}}^\alpha \equiv \hat{\mathbf{a}}(\mathcal{A}^\alpha, \mathcal{A})$ in each node subset. For any two aligned subsets of nodes, $\mathcal{A}^\alpha$ and $\mathcal{A}^\beta$, we can define the *pairwise mismatch* of their patterns

$$M(\hat{\mathbf{a}}^\alpha, \hat{\mathbf{a}}^\beta) = \sum_{i,i'=1}^n [\hat{a}_{ii'}^\alpha (1 - \hat{a}_{ii'}^\beta) + (1 - \hat{a}_{ii'}^\alpha) \hat{a}_{ii'}^\beta]. \quad (4.8)$$

The mismatch is a Hamming distance for aligned patterns. The average $\overline{M}$ of the mismatch over all pairs of aligned patterns is termed the *fuzziness* of the alignment.

Frequently network motifs also have an enhanced number of internal links,^{26,27} providing the possibility of feedback or other faculties not available to tree-like patterns. An ensemble describing p node subsets with correlated patterns $\hat{\mathbf{a}}^1, \dots, \hat{\mathbf{a}}^p$ with an enhanced number of links is given by

$$Q_{\mu,\sigma}(\hat{\mathbf{a}}^1, \dots, \hat{\mathbf{a}}^p) = Z_{\mu,\sigma}^{-1} \prod_{\alpha=1}^p P_0(\hat{\mathbf{a}}^\alpha) \quad (4.9)$$

$$\times \exp \left[-\frac{\mu}{2p} \sum_{\alpha,\beta=1}^p M(\hat{\mathbf{a}}^\alpha, \hat{\mathbf{a}}^\beta) + \sigma \sum_{\alpha=1}^p L(\hat{\mathbf{a}}^\alpha) \right].$$

The parameter $\mu \geq 0$ biases the ensemble (4.9) towards patterns with small mutual mismatches $M(\hat{\mathbf{a}}^\alpha, \hat{\mathbf{a}}^\beta)$.

Given the null model (4.1) and the model (4.9) with correlated patterns, we obtain a log-likelihood score for network motifs

$$S(\mathcal{A}, \mu, \sigma)$$

$$= \log \left(\frac{Q_{\mu,\sigma}(\hat{\mathbf{a}}^1, \dots, \hat{\mathbf{a}}^p)}{P_0(\hat{\mathbf{a}}^1, \dots, \hat{\mathbf{a}}^p)} \right)$$

$$= -\frac{\mu}{2p} \sum_{\alpha,\beta=1}^p M(\hat{\mathbf{a}}^\alpha, \hat{\mathbf{a}}^\beta) + \sigma \sum_{\alpha=1}^p L(\hat{\mathbf{a}}^\alpha) - \log Z_{\mu,\sigma}. \quad (4.10)$$

High-scoring alignments $\mathcal{A}$ indicate *bona fide* network motifs. The first and second terms reward alignments with a small mutual mismatch and a high number of internal links, respectively. The term $\log Z_{\mu,\sigma}$ acts as a threshold assigning a negative score to alignments with too high fuzziness or too few internal links.

Again, both the alignment $\mathcal{A}$ and the scoring parameters μ and σ are *a priori* undetermined. For given scoring parameters, the maximum-score alignment

$$\mathcal{A}^*(\mu, \sigma) = \operatorname{argmax}_{\mathcal{A}} S(\mathcal{A}, \mu, \sigma) \quad (4.11)$$

occurs at some finite value of the number of subgraphs $p^*(\mu, \sigma)$.

The scoring parameters μ and σ can again be determined by maximum likelihood, which corresponds to maximising the score $S(\mathcal{A}^*(\mu, \sigma), \mu, \sigma)$ with respect to the scoring parameters. By differentiating (4.10) with respect to the scoring parameters one finds that at $\mu = \mu^*$ and $\sigma = \sigma^*$ the model (4.9) fits the maximum-score network motifs: the expectation values of the internal number of links and the fuzziness equal the corresponding values of the maximum-score alignment.

4.5.1. Network motifs in regulatory networks

We now apply the scoring function (4.10) to the identification of network motifs in the gene regulatory network of *E. coli*, taken from Ref. 26. A full account and

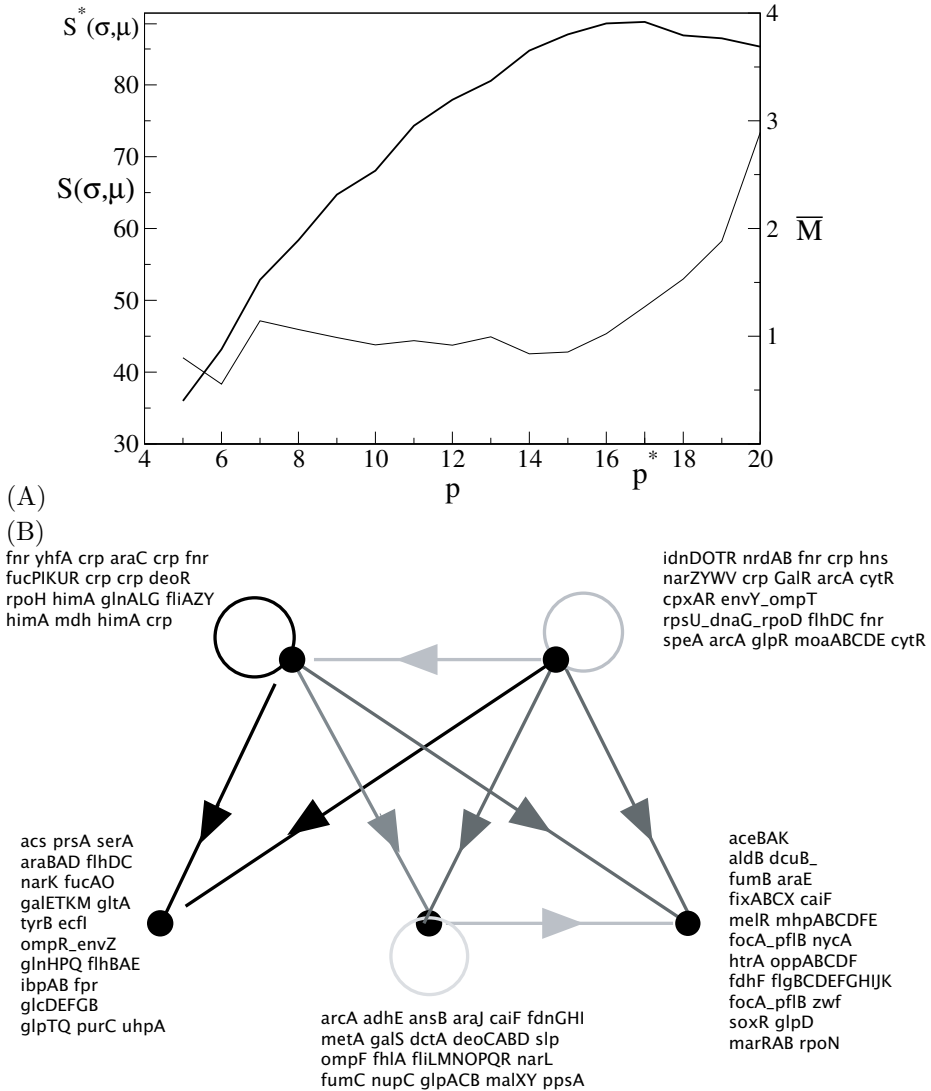


Fig. 4.3. Motifs in the regulatory network of *E. coli*. (A) Score optimisation at fixed scoring parameters $\sigma = 3.8$ and $\mu = 4.0$ for subgraphs of size $n = 5$. The total score S (thick line) and the fuzziness $\bar{M}$ (thin line) are shown for the highest-scoring alignment of p subgraphs, plotted as a function of p . (B) The consensus motif of the optimal alignment, and the identities of the genes involved. The alignment consists of 18 subgraphs sharing at most one node. The five grey values correspond to the consensus motif $\bar{a}$ defined by Eqn. (4.12) in the range 0.1-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8 and 0.8-0.9.

a score-maximisation algorithm are given in Ref. 28. We first investigate the properties of the maximal score alignment at fixed scoring parameters. Fig. 4.3A shows the score S and the fuzziness $\bar{M}$ for the highest-scoring alignment with a

prescribed number p of subgraphs, plotted against p . The fuzziness increases with p and the score reaches its maximum $S^*(\sigma, \mu)$ at some value $p^*(\sigma, \mu)$. For $p < p^*(\sigma, \mu)$ the score is lower since the alignment contains fewer subgraphs, and for $p > p^*(\sigma, \mu)$ it is lower since the subgraphs have higher mutual mismatches.

The optimal scoring parameters μ and σ are again inferred by maximum likelihood. The resulting optimal alignment $\mathcal{A}^* \equiv \mathcal{A}^*(\mu^*, \sigma^*)$ is shown in Fig. 4.3B using the so-called *consensus motif*

$$\bar{\mathbf{a}} = \frac{1}{p} \sum_{\alpha=1}^p \hat{\mathbf{a}}^\alpha(\mathcal{A}^*) . \quad (4.12)$$

The consensus motif is a *probabilistic pattern*; the entry $\bar{a}$ denotes the probability that a given binary link is present in the aligned subgraphs. The motif shown in Fig. 4.3B consists of 2 + 3 nodes forming an input and an output layer, with links largely going from the input to the output layer. Most genes in the input layer code for transcription factors or are involved in signalling pathways. The output layer mainly consists of genes coding for enzymes.

4.6. Cross-Species Analysis of Networks

The motifs discussed above show correlation without sharing a common evolutionary history. Larger functional units may be distinguished by their evolutionary conservation. Thus, we expect parts of the network to maintain their topology and to form a conserved core, while other parts show a more rapid turnover of both nodes and interactions, see Fig. 4.1C. This conservation can be detected as topological correlation across species.

We assume that organisms evolve independently after speciation, leading to divergence in their network links as well as in the overall similarity of the nucleotide sequences, the structure of proteins, and the biochemical role of a metabolite. The relationship between link and node similarity is non-trivial: genes may retain their function and their interactions with other genes despite considerable sequence divergence. On the other hand, the change of a few nucleotides can create or destroy a binding site, implying that genes with high overall sequence similarity may have entirely different interactions. Hence, cross-species analysis has to take into account information from both links and nodes.

A log-likelihood score assessing the link statistics of node subsets in network A and in network B follows directly from Eqn. (4.10). This *link score* is given by

$$\begin{aligned} S^\ell(\mathcal{A}, \mu, \sigma_A, \sigma_B) &= -\mu M(\hat{\mathbf{a}}, \hat{\mathbf{b}}) \\ &+ \sigma \left(L(\hat{\mathbf{a}}) + L(\hat{\mathbf{b}}) \right) - \log Z(\mu, \sigma_A, \sigma_B) . \end{aligned} \quad (4.13)$$

To assess the similarity of nodes, we consider a measure θ_{ij} , which describes the similarity of node i in network A and node j in network B . The node similarity measure may be a percentage sequence identity, or a distance measure of protein

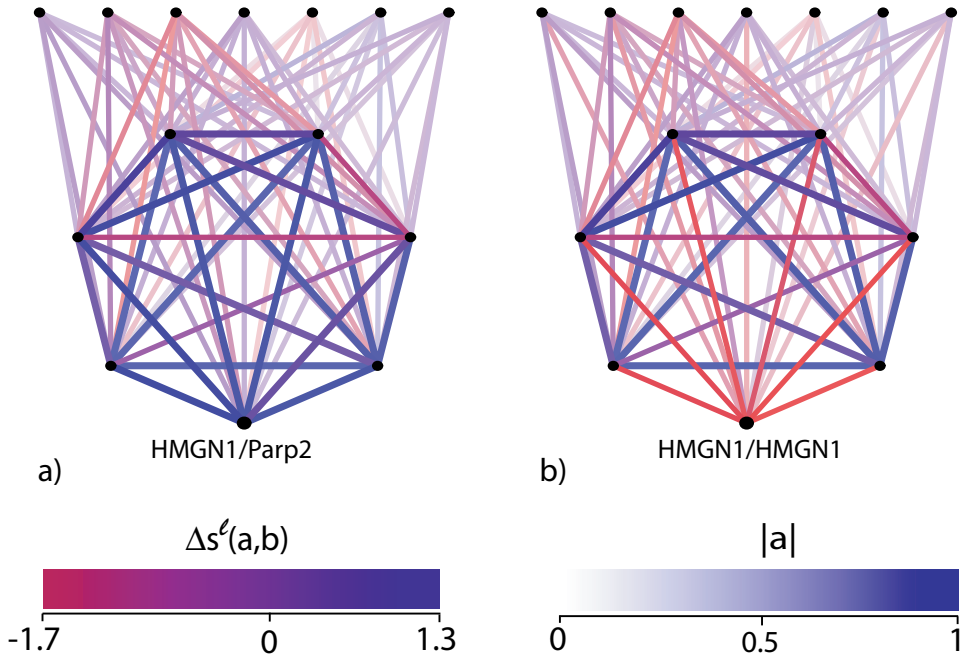


Fig. 4.4. Cross-species network alignment shows conservation of gene clusters. (A) Seven genes from a cluster of co-expressed genes (circle) together with seven random genes outside the cluster (straight line). Each node represents a pair of aligned genes in human and mouse. The intensity of a link encodes the correlation coefficient a of gene expression patterns in human, see text. The colour indicates the evolutionary conservation of a link, with blue hues indicating strong conservation. The conservation is quantified by the excess link score contribution, Δs^ℓ , defined as the link score minus the average link score of links with the same correlation value. (B) The same cluster, but with human-HMGN1 ‘falsely’ aligned to its orthologue mouse-HMGN1, with the red links showing the poor expression overlap of this pair of genes.

structures. The information on node similarity can be incorporated into the alignment score by contrasting a null model with a model describing a statistic where node similarity is correlated with the alignment. To construct the null model, we assume that node similarities θ_{ij} for different node pairs i, j are identically and independently distributed and denote their distribution by $p_0^n(\theta_{ij})$. The model describing cross-species correlations has to take into account that the distribution of node similarities between aligned pairs of nodes follows a different statistic (typically generating higher values of θ), denoted by $q_1^n(\theta)$. The distribution of pairwise similarity coefficients between one aligned node and nodes other than its alignment partner is denoted by $q_2^n(\theta)$. Assuming that the statistics of links and nodes similarities are uncorrelated for a given alignment, a simple calculation analogous to Eqn. (4.4) yields the log-likelihood score

$$S(\mathcal{A}) = S^\ell(\mathcal{A}) + S^n(\mathcal{A}) , \quad (4.14)$$

with the information from node similarity contributing a *node score*

$$S^n(\mathcal{A}) = \sum_{i \in \mathcal{A}} s_1^n(\theta_{ii}) + \sum_{\substack{i \in \mathcal{A}, j \neq i \\ j \in \mathcal{B}, i \notin \mathcal{A}}} s_2^n(\theta_{ij}) \quad (4.15)$$

and $s_1^n(\theta) \equiv \log(q_1^n(\theta)/p_0^n(\theta))$ and $s_2^n(\theta) \equiv \log(q_2^n(\theta)/p_0^n(\theta))$. The number of nodes in the two networks can be different from each other. Nodes may lack an alignment partner due to node loss in one lineage, or because of a high degree of link dynamics.

The scoring parameters entering Eqn. (4.14) need to be determined from the data. Provided there are not too many scoring parameters, this can again be done by maximum likelihood as outlined in the preceding sections. Particular examples are networks with binary links and coarse-grained measures of sequence similarity. (As an extreme case, node similarity may be considered a binary variable, when nodes either have significant similarity or not. Then the ensembles describing the node statistics are each described by a single variable, see Ref. 29 for details.)

4.6.1. Alignment of co-expression networks

We now compare co-expression networks of *H. sapiens* and *M. musculus*. In co-expression networks, the weighted link $a_{ii'} \in [-1, 1]$ between a pair of genes i, j is given by the correlation coefficient of their gene expression profiles measured on a microarray chip. Genes which tend to be expressed under similar conditions thus have positive links. The score (4.13) can easily be generalised to weighted interactions, see Ref. 29.

The data of Su *et al.*³⁰ was used to construct networks of ~ 2000 housekeeping genes. Human-mouse orthologues were taken from the Ensembl database.²³ Details on the algorithm to maximise the score (4.13) are given in Ref. 29.

We focus on strongly conserved parts of the two networks. Figure 4.4 shows a cluster of co-expressed genes which is highly conserved between human and mouse (link conservation is shown in blue, changes between the links in red).

With one exception, the aligned gene pairs in this cluster have significant sequence similarity and are thought to be orthologues, stemming from a common ancestral gene. The exception is the aligned gene pair human-HMGN1/mouse-Parp2. These genes are aligned due to their matching links, quantified by a high contribution to the link score (4.13) of $S^\ell = 25.1$. The ‘false’ alignment human-HMGN1/mouse-HMGN1 respects sequence similarity but produces a link mismatch ($S^\ell = -12.4$); see Fig. 4.4B. Human-HMGN1 is known to be involved in chromatin modulation and acts as a transcription factor. The network alignment predicts a similar role of Parp2 in mouse, which is distinct from its known function in the poly(ADP-ribosyl)ation of nuclear proteins. The prediction is compatible with ex-

periments on the effect of Parp-inhibition, which suggest that Parp genes in mouse play a role in chromatin modification during development.³¹

4.7. Towards an Evolutionary Theory

Different parts of biological networks have different functions. Here we have applied a statistical approach to the detection of network clusters, network motifs and cross-species correlations. But the detection of deviations from a global background statistics has a wider perspective, which includes the connection between different type of networks, the link between network topology and the underlying sequence, and spatiotemporal changes of biological networks. From an evolutionary point of view, these deviations are created and maintained by selection pressures which are both non-homogeneous and correlated across the network. A quantitative theory of biological networks will thus require a synthesis of network statistics and population genetics, a largely outstanding task to date. Here we give a brief outlook on some of the challenges ahead.

4.7.1. *Genetic interactions between different links*

Biological function is typically tied to modules consisting of several nodes and links. As a result, there are correlations between links across different species: a species with a certain function will tend to have all links associated with the specific function, a species lacking the function will tend to have none of the corresponding links. The network motifs discussed above are only a special case of this phenomenon. With data on biological networks becoming available for an increasing number of species, it will become feasible to infer these correlations and the corresponding functional modules from data. Scoring functions constructed to detect genetic interactions in multiple alignments will play an important role in this undertaking.

4.7.2. *Gene duplications*

Following the duplication of a gene, the daughter genes have the same function and same interactions with other genes. Independent evolution of the two genes may lead to the non-functionalisation and even the loss of one of the duplicates, or to sub-functionalisation, with different functional roles being divided between the two copies.³² Tracing the dynamics of gene duplication at the level of interaction networks gives insight into the evolutionary dynamics of networks.^{20,33} Scoring for jointly conserved subgroups of links can be used to identify the different functional modules a gene is involved in. This can be done both at the level of single species, as well as in a cross-species analysis, where gene duplications introduce one-to-many and many-to-many alignments.

4.7.3. *Neutral and selective dynamics*

Biological networks show a great deal of plasticity, since the same biological function can be carried out by different networks (see e.g. Ref. 34). This flexibility leads to neutral evolution as a population explores the space of networks corresponding to a given function. On the other hand, networks may change as a new functionality is acquired, or because of changing environmental conditions. Disentangling neutral moves and changes under selection is possible by contrasting inter-species variability with intra-species variability.³⁵ Inferring the modes of network evolution and the relative weights of neutral and selective dynamics remains an outstanding challenge for experiment and theory.

Acknowledgements

This work was supported through DFG grants SFB/TR 12, SFB 680 and BE 2478/2-1. We thank David Arnosti, Daniel Barker, Leonid Mirny and Nina White for the discussions.

Appendix: Bayesian Analysis of Network Data

The detection of deviations from a null model can be formulated as a problem of deciding between alternative hypotheses. The first hypothesis is that a given node subset follows the statistic of the null model. The alternative hypothesis is that the node subset follows a statistic different from the null model. This statistic is called the Q -model.

The choice between these two alternatives can be formulated probabilistically by considering the *posterior probability* $P(Q|\hat{\mathbf{a}}, \mathcal{A})$. It describes the probability that the node subset(s) specified by $\mathcal{A}$ follow the Q -model (hypothesis Q), rather than the null model (null-hypothesis P_0). Denoting any prior knowledge we may have about the probability with which the two alternatives occur by $P(Q)$ and $P(P_0)$, respectively, one may use Bayes' theorem to find

$$\begin{aligned} P(Q|\hat{\mathbf{a}}, \mathcal{A}) &= \frac{P(\hat{\mathbf{a}}|Q, \mathcal{A})P(Q)}{P(\hat{\mathbf{a}}|\mathcal{A})} \\ &= \frac{P(\hat{\mathbf{a}}|Q, \mathcal{A})P(Q)}{P(\hat{\mathbf{a}}|P_0, \mathcal{A})P(P_0) + P(\hat{\mathbf{a}}|Q, \mathcal{A})P(Q)} \\ &= \frac{e^{S'(\mathcal{A})}}{1 + e^{S'(\mathcal{A})}} . \end{aligned} \tag{4.16}$$

$P(\hat{\mathbf{a}}|Q, \mathcal{A})$ gives the probability of generating patterns $\hat{\mathbf{a}}$ under the Q -model (given, for instance, by Eqn. (4.3) or by Eqn. (4.9)). $P(\hat{\mathbf{a}}|P_0, \mathcal{A})$ gives the probability of generating the same pattern under the null model (4.1). The posterior probability

is thus a monotonously increasing function of the *log-likelihood score* given by

$$\begin{aligned} S'(\mathcal{A}) &= \log \left(\frac{P(\hat{\mathbf{a}}|Q, \mathcal{A})}{P(\hat{\mathbf{a}}|P_0, \mathcal{A})} \right) + \log \left(\frac{P(Q)}{P(P_0)} \right) \\ &= S(\mathcal{A}) + \text{const.} \end{aligned} \quad (4.17)$$

Hence the score $S(\mathcal{A})$ defined in Eqn. (4.4) has a sound theoretical foundation: it is a measure of the posterior probability that the node subset specified by $\mathcal{A}$ follows the Q -model rather than the null model.

This simple picture needs to be extended when the parameters m of the Q -model and the alignment $\mathcal{A}$ are unknown and are considered ‘hidden’ variables to be determined from the data. We construct a model of the entire network with adjacency matrix $\mathbf{a}$, with pattern $\hat{\mathbf{a}}(\mathcal{A})$ following the Q -model and the remainder of the network following the null model

$$P(\mathbf{a}|\mathcal{A}, m) = Q(\hat{\mathbf{a}}|\mathcal{A}, m)P_0(\tilde{\mathbf{a}}|\mathcal{A}) . \quad (4.18)$$

The matrix of links between nodes which are not both part of $\mathcal{A}$ is denoted by $\tilde{\mathbf{a}}$. Using Bayes’ theorem one can write the posterior probability of $\mathcal{A}$ and m , i.e. the conditional probability of the hidden variables, in the form

$$P(\mathcal{A}, m|\mathbf{a}) = \frac{Q(\mathbf{a}|\mathcal{A}, m)P(\mathcal{A}, m)}{\sum_{\mathcal{A}, m} Q(\mathbf{a}|\mathcal{A}, m)P(\mathcal{A}, m)} . \quad (4.19)$$

We assume the prior probability $P(\mathcal{A}, m)$ to be flat. Dropping the terms independent of $\mathcal{A}$ and m , the optimal alignment $\mathcal{A}^*$ is obtained by maximising the posterior probability $Q(\mathcal{A}|\mathbf{a}) \sim \sum_m Q(\mathbf{a}|\mathcal{A}, m)$ with respect to $\mathcal{A}$ and similarly the optimal scoring parameters m^* by maximising $Q(m|\mathbf{a}) \sim \sum_{\mathcal{A}} Q(\mathbf{a}|\mathcal{A}, m)$ with respect to m . In the so-called Viterbi approximation, $\mathcal{A}^*$ and m^* are inferred by jointly maximising $Q(\mathbf{a}, \mathbf{b}, \Theta|\mathcal{A}, m)$ with respect to $\mathcal{A}$ and m . Assuming the sum $\sum_{\mathcal{A}, m} Q(\mathbf{a}|\mathcal{A}, m)$ can be split into the term stemming from $\mathcal{A}^*, m^*$ and a remainder $\sum_{\mathcal{A} \neq \mathcal{A}^*, m \neq m^*} Q(\mathbf{a}|\mathcal{A}, m) \sim P_0(\mathbf{a})$, the posterior probability (4.19) can again be written in the form of Eqn. (4.17). In this approximation, the maximum-score alignment and the optimal scoring parameters are determined by the maximum of the log-likelihood score (4.4) over the alignments and over the scoring parameters.

References

1. L. D. Stein. Human genome: End of the beginning. *Nature*, 431:915 – 916, 2004.
2. J.-M. Claverie. What if there are only 30,000 human genes? *Science*, 291(5507):1255–1257, 2001.
3. euGenes-database. <http://eugenesis.org/all/homologies/hgsummary-2002.html>.
4. M.C. King and A.C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188:107–166, 1975.
5. D. Tautz. Evolution of transcriptional regulation. *Current Opinion in Genetics & Development*, 10:575–579, 2000.
6. G.A. Wray. Transcriptional regulation and the evolution of development. *Int J Dev Biol*, 47(7-8):675–684, 2003.

7. J. Berg, S. Willmann, and M. Lässig. Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology*, 4(1):42, 2004.
8. M.S. Gelfand. Evolution of transcriptional regulatory networks in microbial genomes. *Curr Opin Struct Biol*, 16(3):420–429, 2006.
9. R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis*. CUP, Cambridge, UK, 1998.
10. P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403:623–627, 2000.
11. S. Li, C. M. Armstrong, N. Bertin, Hui Ge, S. Milstein, et al. A map of the interactome network of the metazoan *C. elegans*. *Science*, 303(5657):540–543, Jan 2004.
12. L. Giot, J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 302(5651):1727–1736, 2003.
13. J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.
14. Yingming Zhao, T. W. Muir, S. B.H. Kent, E. Tischer, J. M. Scardina, and B. T. Chait. Mapping protein–protein interactions by affinity-directed mass spectrometry. *PNAS*, 93(9):4020–4024, 1996.
15. C. E Horak and M. Snyder. ChIP-chip: a genomic approach for identifying transcription factor binding sites. *Methods Enzymol*, 350:469–483, 2002.
16. L. M. Smoot, J. C. Smoot, H. Smidt, P. A. Noble, M. Konneke, et al. DNA microarrays as salivary diagnostic tools for characterizing the oral cavity’s microbial community. *Adv Dent Res*, 18(1):6–11, 2005.
17. C. Stremmel, A. Wein, W. Hohenberger, and B. Reingruber. DNA microarrays: a new diagnostic tool and its implications in colorectal cancer. *Int J Colorectal Dis*, 17(3):131–136, 2002.
18. A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
19. A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *Complexus*, 1:38–44, 2003.
20. J. Berg, M. Lässig, and A. Wagner. Structure and evolution of protein interaction networks: A statistical model for link dynamics and gene duplications. *BMC Evolutionary Biology*, 4:51, 2004.
21. S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon. Subgraphs in random networks. *Phys. Rev.*, 68:026127, 2003.
22. U. Einav, Y. Tabach, G. Getz, A. Yitzhaky, U. Ozbek, et al. Gene expression analysis reveals a strong signature of an interferon-induced pathway in childhood lymphoblastic leukemia as well as in breast and ovarian cancer. *Oncogene*, 24(42):6367–6375, 2005.
23. T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, et al. Ensembl 2005. *Nucleic Acids Res.*, 33:D447–D453, 2005.
24. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29, 2000.
25. P. A. Padilla, E. K. Fuge, M. E. Crawford, A. Errett, and M. Werner-Washburne. The highly conserved, coregulated SNO and SNZ gene families in *Saccharomyces cerevisiae* respond to nutrient limitation. *J. Bacteriol.*, 180:5718–5726, 1998.
26. S. Shen Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31:64–68, 2002.
27. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.

28. J. Berg and M. Lässig. Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. USA*, 101(41):14689–14694, 2004.
29. J. Berg and M. Lässig. Cross-species analysis of biological networks by Bayesian alignment. *Proc. Natl. Acad. Sci. USA*, in press, 2006.
30. A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, 2004.
31. T. Imamura, T. M. Anh, C. Thenevin, and A. Paldi. Essential role for poly (adp-ribosyl)ation in mouse preimplantation development. *BMC Molecular Biology*, 5:4, 2004.
32. M. Lynch, M. O’Hely, B. Walsh, and A. Force. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159:1789–1804, 2001.
33. W.-Y. Chung, R. Albert, I. Albert, A. Nekrutenko, and K.D. Makova. Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC Bioinformatics*, 7:46, 2006.
34. A. Tanay, A. Regev, and R. Shamir. Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. USA*, 2005.
35. J. H. McDonald and M. Kreitman. Adaptive protein evolution at Adh locus in *Drosophila*. *Nature*, 351:652–654, 1991.

This page intentionally left blank

Chapter 5

Network Concepts and Epidemiological Models

ROWLAND R. KAO¹ AND ISTVAN Z. KISS²

¹*Institute of Comparative Medicine, University of Glasgow*

²*Department of Mathematics, University of Sussex*

r.kao@vet.gla.ac.uk, I.Z.Kiss@sussex.ac.uk

Mathematical approaches to study the dynamics of infectious diseases go back many years. They have primarily built on differential equations assuming individuals are mixing randomly with no population structure. In contrast, under the network paradigm, a population is a network allowing individuals to interact with their neighbours in the network, i.e. the links between individuals represent potential transmissions of disease. In this chapter we review current development in network epidemiology and relate it to the classical modelling and discuss different types of network structures such as small-world and scale-free networks.

5.1. Introduction

The development of a mathematical approach to studying the population dynamics of infectious diseases can be traced to the work of Sir Ronald Ross, a polymath who won a Nobel Prize in medicine for identifying the role of the *Anopheles* mosquito in the transmission of malaria. Ross' remarkable body of work consisted of experiments, field investigations and the development of a theoretical framework based on a mathematical description of the malaria host-parasite system.⁵⁷ Ross' mathematical description was later extended and generalised by Kermack and McKendrick,⁴¹ whose work forms the basis for the *SIR* differential equation model which lies at the heart of modern quantitative epidemiology. The Kermack–McKendrick model was originally developed in the context of a set of integro-differential equations, using an infection-structured formulation allowing for flexible interpretation of the rates of transmission over the infection lifetime. The modernly accepted Kermack–McKendrick model makes the simplification of assuming a single exponentially distributed infectious stage, with all infected individuals being equally infectious. With this assumption, the system takes the form of a *compartmental model*, here a set of three ordinary differential equations to be integrated over time:

$$\begin{aligned}\frac{dS}{dt} &= -\beta IS \\ \frac{dI}{dt} &= \beta IS - \gamma I\end{aligned}\tag{5.1}$$

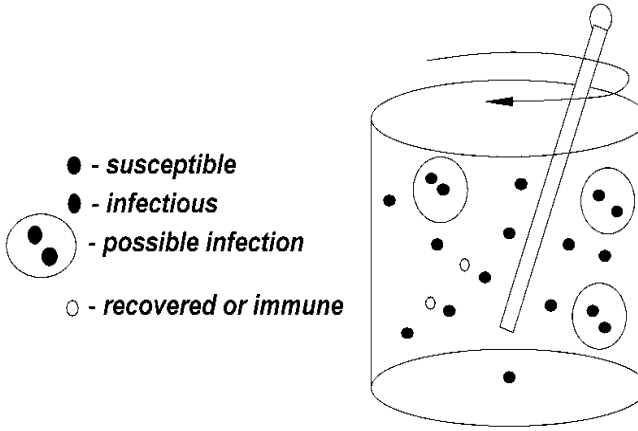


Fig. 5.1. Homogeneous random mixing can be viewed as a ‘well-stirred system’, where infected individuals are equally likely to interact with any other member of the population.

$$\frac{dR}{dt} = \gamma I$$

$$S + I + R = N.$$

In the system of Eqn. 5.1, the compartments are the number of susceptible individuals S , the number of infected I and the number of removed R (usually considered to be recovered and immune, though other interpretations of this state are possible). The parameter β is the rate per infected individual at which infections occur, while γ is the rate at which infected individuals are removed.

Some of the key principles that have guided much of mathematical epidemiology over the last century are apparent in this simple formulation. First, interest in the field has concentrated on the non-linear interactions between a host population and a pathogen that exploits it. Second, it is assumed that, for the purposes of gaining insight into the dynamics of disease spread at the population level, individuals can be treated as indistinguishable except for their disease state. Third, interactions between members of the population are considered to occur at random, with equal probability that any member will interact with any other element of the system (Fig. 5.1). Finally, the model operates in continuous space, time and population-space.

In contrast, under the network paradigm of disease spread, a population is a network (in mathematical theory, *graph*) that consists of a set of nodes (*vertices*) representing epidemiological units at a given scale (e.g. individuals, towns, cities, farms or wildlife communities). Each node i is connected to other nodes in the network by k_i links (*edges*), this defining the degree of the node. The links usually represent potentially infectious contacts. For example, for sexually transmitted infections or STIs, links may be sexual acts or sexual partners, while for diseases transmitted within a hospital, links may represent contacts occurring through room-

and ward-sharing. The links may be directed or undirected and the probability of transmission across links weighted or unweighted (i.e. any infected node has the same probability of infecting any susceptible node if they are directly connected to each other). Probabilities of transmission are usually independent (i.e. if a node is connected to two infected nodes, each of which can infect with probability $\bar{p}$, the probability of becoming infected is $(1 - (1 - \bar{p})^2)$). In directed networks (e.g. where one individual can infect another but not necessarily *vice versa*), links are distinguished as being in- or out-links, with nodes having in- and out-degrees. In most examples, $\langle k \rangle \ll N$, where $\langle k \rangle$ is the average node degree and N the population size. Nodes typically possess one of a limited number of states (e.g. susceptible, infected or removed as in the Kermack–McKendrick model). *Mean-field* models such as that described by Eqn. (5.1) are similar to maximally connected network models – i.e. where every individual in the population is connected to any other individual and $\langle k \rangle = k_i = N - 1$ for all nodes i . In this sense, network models can be viewed as a generalisation of mean-field models. However, mean-field and network models differ in terms of the philosophy behind their representations. Mean-field models often do have population structure, but with this structure being imposed on the population, rather than being generated from individual properties. In contrast, from the network perspective, each node only has information about a limited subset of the entire population. Links are generated from this ‘local neighbourhood’ that defines the social network. Thus population structure is defined by these individual properties, and the network model displays corresponding *emergent behaviour* in a way that the Kermack–McKendrick model does not. Of course, both pattern (population structure) and process (the nature of the interactions highlighted in mean-field models) are important in determining how epidemics are spread. That most work has previously concentrated on the dynamics amongst simplified compartments is at least partially pragmatic – observational data on overall disease incidence and detailed data describing the time course of individual infection states have historically been more available than meaningful population contact structure data, particularly for humans. For example, one of the most detailed and successful models of disease transmission in structured large human populations is the description of measles outbreaks in post-WWII Britain^{11,28} which includes comprehensive measles incidence reports, but where location is only specified to the level of city or town. Potentially infectious connections between cities are handled abstractly. The development of the field has also benefited from the rich literature of dynamical systems and the development of analogous models in chemical kinetics, reflected in the early appellation of *mass-action* dynamics when referring to what is now commonly known as *density dependent* contact.* Despite this emphasis, the importance of contact heterogeneity has of course been recognised. An important point that will

*Note that there has been some confusion on this, see De Jong M.C.M., Bouma A., Diekmann O., Heesterbeek H. (2002) Modelling transmission: mass action and beyond. *Trends in Ecology and Evolution* 17: 64

be developed here is that many of the ideas explored in social network approaches have been previously explored using other approaches, though in many ways the social network paradigm has often proved to be more natural, and provided insights that would not so easily be explored in other contexts.

One way of looking at social network analysis is as a ‘middle way’ between the highly simplified contact structures typified by Eqn. (5.1), and extremely complex simulations which, like social networks, are individual-based but typically involve many parameters.^{22,24} Another interpretation is that, while ODE models concentrate on the temporal dynamics of disease transmission at the expense of simplifying the spatial or contact structure, network analyses at their simplest only consider abstract temporal dynamics, not allowing for varying infectiousness over time, for example. Whatever the philosophical interpretation, network models retain some of the simplicity and analytical tractability of the former, while introducing in a natural way the study of complex contact structures. Especially as high performance computing devices have become common, detailed simulations have become increasingly popular and useful research tools. Nevertheless the analysis of simplified structures such as social networks is vital for gaining insight into how heterogeneity in the contacts amongst individuals can contribute to disease spread and its control. Here, we concentrate on the development of two critical ideas in the development of social network theory (small-world networks and scale-free distributions) and emphasise two themes – what the social network approach has added to the already rich literature of mathematical epidemiology, and how consideration of epidemic dynamics changes the way we perceive network structure.

5.2. Simple Epidemiological Models

5.2.1. *Introducing R_0*

For compartmental models of disease spread, the stability of the disease-free state is determined by the basic reproduction number, the central quantity of modern theoretical epidemiology,^{5,16} generally denoted by the symbol R_0 . The ‘simple’, commonly accepted biological definition of R_0 is generally stated as ‘the number of new infections generated by a single infected individual introduced into a wholly susceptible, homogeneously mixed population at equilibrium’. For the system of Eqn. (5.1), it is easy to show that this definition is equivalent to:

$$R_0 = \frac{\beta N}{\gamma}. \quad (5.2)$$

For simple systems, if $R_0 < 1$, then the disease-free state is globally asymptotically stable (but see section below). Each person who contracts the disease will on average infect fewer than one person before dying or recovering, so the outbreak itself will die out (i.e. $dI/dt < 0$). When $R_0 > 1$, each person who becomes infected will infect on average more than one person, so the epidemic will spread ($dI/dt > 0$). While

this definition is intuitive, conceptual problems immediately arise. For example, can one define a ‘typical’ infected individual? At what stage of the infection process is the infected individual introduced? What if there are distinct subpopulations or population structures? Is R_0 then a meaningful concept? Considerable attention has been devoted to these questions.^{16,30,56,60} In particular, most network models with their complex structure do not lend themselves to such simple definitions, and the relationship between R_0 and the network representation is further discussed below.

5.2.2. *Density vs. frequency dependent contact*

A connection between Eqn. (5.1) and network models can be established by a closer examination of the contact structure implicit in the nonlinear term βSI , which can be written more generally if we replace the expression βSI

$$\beta C(N) I \times \frac{S}{N}$$

(see for example Ref. 55), where each individual has $C(N)$ potential infectious contacts, a number which is dependent on the total population N .[†] The region in parameter space where $R_0 < 1$ then defines a globally stable disease-free state if $dC/dN \geq 0$ (usually, $d^2C/dN^2 \leq 0$ but this is not required), and that none of $C(N)$, β or γ are functions of I . In particular if $dC/dI > 0$, $d\beta/dI > 0$, or $d\gamma/dI > 0$, global stability is lost. There are various ways for these to occur. For example, if removal of infected individuals requires the availability of limited resources, $d\gamma/dI > 0$ (e.g. foot-and-mouth disease in the UK in 2001, see Ref. 29) or one may have $dC/dI > 0$ if contacts are increased by otherwise sedentary individuals attempting to flee an epidemic, as may have occurred during the Black Death in 14th century Europe. Each infected individual has a probability S/N per contact of interacting with a susceptible individual. For density dependent contact, $C(N) = N$ and the form of Eqn. (5.1) is obtained. For *frequency dependent* contact, $C(N) = \kappa$, a constant. In this case, the rate that new infections appear is $\beta SI\kappa/N$, and $R_0 = \beta\kappa/\gamma$. A critical difference between the two is that in the density dependent case, thinning of the total population reduces N and therefore the value of R_0 , while with frequency dependence the reduction in population density or size has no effect on R_0 . Frequency dependent models correspond to network models in that the number of contacts (links) does not scale with population size. However, frequency dependent models have only a fixed number of contacts per individual (thus a degree distribution with zero variance) and it is not specified with whom these contacts are made. Thus the two are only equivalent in the case of a network with links that switch to random nodes at an infinite rate.⁵³ Most importantly any infected individual is still assumed to have κ outward potentially infectious

[†]We note that this it is sometimes more important to consider population density rather than total population, however we will consider dynamics that depend on population size.

contacts, while in static network models one of the links is ‘used up’ because the node was infected through one of its existing links.¹⁵

5.3. Some Definitions and Their Application to Poisson Random Networks

Network structure enriches our understanding of how diseases might spread through a population. As previously noted, in network models individuals can no longer be assumed to be in potentially infectious contact with all members of the population. Thus the degree distribution, average path length, path length distribution and the diameter of the network are quantitative measures that offer insight into how well connected a network is, and therefore the risk that large proportions of the population become infected or that particular subgroups are more likely to become infected.

The degree distribution $p(k)$ gives the probability that a randomly selected node has exactly k links. The average number of connections per node is given by $\langle k \rangle = \sum_l lp(l)$. Epidemiologically the degree of a node gives the maximum number of nodes that it could infect. Of course, as $\langle k \rangle \ll N$, only a few nodes are likely to be infected by any given node. Thus considering the set of nodes that can form a series of connections linking two arbitrary members of the population is important. The path length between two nodes of the network is defined as the minimum number of links needed to connect them (when two nodes are disconnected the path length is considered to be infinite) and the spread in all possible shortest path lengths is captured by the path length distribution. The diameter of the network is the maximum shortest path length between all the possible pairs of the network nodes.

In a Poisson random network (originally studied by Erdős and Rényi²¹), nodes are connected by links, these chosen randomly from the $N(N-1)/2$ possible links. An equivalent definition is the binomial model, where every possible pair out of the nodes is connected with probability $\tilde{p}$. The average number of connections per node is $\langle k \rangle = \tilde{p}(N-1)$ and the degree distribution is given by

$$P(k) = \binom{N-1}{k} \tilde{p}^k (1-\tilde{p})^{(N-1)-k} \cong \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} \quad (5.3)$$

where the second equality holds when $N \rightarrow \infty$; this motivates its name of Poisson random graph (or network). When $\tilde{p}$ is sufficiently large, random networks tend to have relatively small diameters. In a Poisson random network the number of nodes at a distance l from a given node is well approximated by $\langle k \rangle^l$.¹³ When the whole network is captured starting from a given node, $\langle k \rangle^l \cong N$ and l approaches the network diameter d . Hence, d depends only logarithmically on the number of nodes, and the average path length is also expected to only scale slowly with increasing population size, i.e. $\langle l_{rand} \rangle \propto \ln(N)/\ln(\langle k \rangle)$, with a correspondingly small diameter.

5.4. Networks With Localisation of Contacts: Small Worlds, Clustering, Pairwise Approximations and Moment Closure

5.4.1. *Small worlds*

A contact network with a small diameter such as those found in Poisson networks supports epidemics that, within relatively few generations of infection, spread broadly throughout the network. Thus even for a disease with low probability of transmission and where the disease has been identified within a few generations of infection after its introduction, it would be difficult to identify and isolate subgroups of individuals who are at higher risk of becoming infected. Empirical measurements confirm that many real-world networks have small average path lengths very similar to that of Poisson random networks, but are characterised by greater localisation of connections – i.e. the tendency for links to occur with greater probability than average amongst subgroups of nodes. Localisation is exemplified by lattice models where nodes are positioned on a regular grid of locations and neighbouring individuals are connected. Such lattice models/networks exhibit homogeneous contact but have much longer average path lengths and diameters than Poisson networks. A model that has both properties of localisation and small average path length is the famous *small-world model* of Watts and Strogatz.⁶² They proposed a one-parameter model that interpolates between a regular lattice model and Poisson random graph. Their model starts with a ring lattice with N nodes where each node is connected to an arbitrary fixed number K of its closest neighbours. Two types of small-world networks have commonly been studied. In the original version, a random rewiring of all links is carried out with probability q . A variant with similar properties does not rewire, but adds long-range links randomly, with probability q to generate the same number of long-range links as in the original model (Fig. 5.2). Both approaches produce on average $qKN/2$ long-range links (or more correctly, links that connect nodes at random). As the latter approach simplifies some calculations but has the same key properties as the original model, it will be referred to later in the chapter. For a broad range of q , the small-world model generates networks with the average path length very close to that observed in Poisson random graphs yet with higher localisation. This model is motivated by social structures where most individuals belong to localised communities composed of work colleagues, neighbours or people sharing similar interests. However, some individuals also have connections with individuals that belong to other localised communities, such as relatives living considerable distances away (and thus likely to belong to distant social communities as well) and old acquaintances. The smaller average path length driven by the limited number of long-range connections (shortcuts) makes the network more connected with fewer edges needed to connect any two nodes. A smaller average path length also means a smaller number of infectious generations with a shorter epidemic time scale, and a lower threshold for a large epidemic. The critical idea put forward by

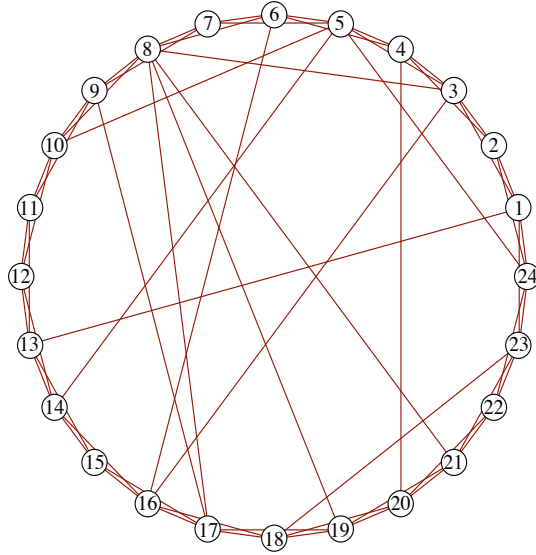


Fig. 5.2. An example of a small-world network, with each node connected locally to its four nearest neighbours.

this model is that relatively few ‘long-distance’ connections are necessary for the transmission and persistence of disease. This has long been established, for example within the metapopulation paradigm developed in the 1960s⁴⁶ where occasional migration between habitat patches was invoked to explain the persistence of species that would otherwise go extinct – in the case of epidemiology, the metapopulation is the pathogen operating on the host (or communities of hosts), which represent the habitat patches, such as the cities and towns in the previously mentioned measles models.^{11,28} Where the model of Watts and Strogatz differed, however, was showing in an elegantly simple model, and in a quantifiable way, how simple couplings defined only as a property of individuals could be weak, yet produce dramatic effects in communities.

5.4.2. *Moment closure*

The small-world model is a very specific, illustrative example of a highly clustered network. More generally, in most populations there are subgroups or communities of individuals that are more likely to be associated with each other, and there is an extensive literature devoted to identifying network-based measures of community (for a review, see Danon *et al.*¹⁴). One measure of localisation is the clustering coefficient, which can be quantified as $c = \frac{3 \times \text{triangles}}{\text{triples}}$, where a triangle is defined by a set of three nodes X , Y and Z in a triplet, where X is connected to Y which is connected to Z , and X is also connected to Z . Thus clustering expresses the

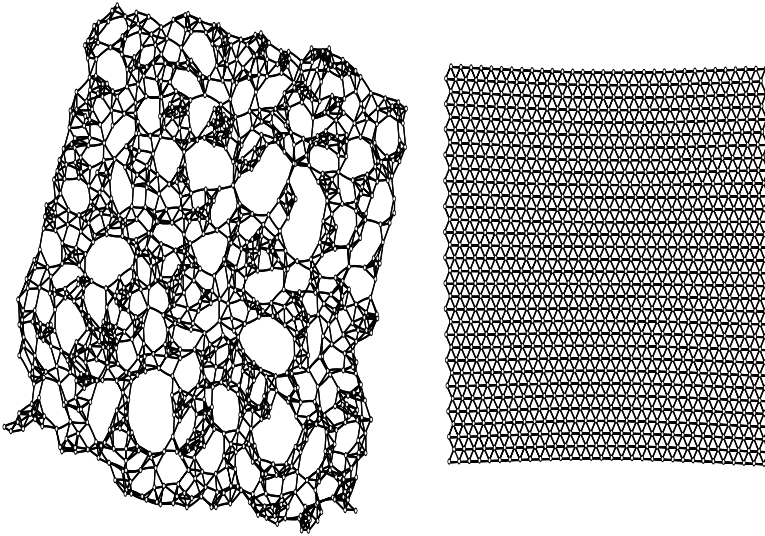


Fig. 5.3. Two social networks with fixed degree distribution $k_i = \langle k \rangle = 6$ and clustering coefficients $c = 0.4$. The network on the left is generated using the Keeling model (1999), the other on the right is a triangular lattice.

probability of two friends of any one individual being themselves friends of each other. This definition is not unique; for example, clustering can also be computed by averaging the clustering coefficients of individual nodes $c_i = \frac{E_i}{k_i(k_i-1)/2}$, which represents the ratio between the number of links E_i present amongst the neighbours of a node and the possible maximum number of such links. In Poisson random networks the inherent clustering $c = \langle k \rangle / (N - 1)$ is small and in the limit of infinite populations, zero. Clustered networks can be generated by randomly distributing individuals/nodes in a given n -dimensional space (e.g. a specified two-dimensional surface) and assuming that the probability of a connection between two individuals is a function of their distance. By choosing an appropriate function the average degree and clustering can be varied. Note that clustering does not uniquely define a network. For example, an infinite number of networks can be generated with zero clustering, and even with nearly identical clustering coefficients, two networks can be quite dissimilar. In Fig. 5.3 a triangular lattice is compared to a network with effectively the same clustering coefficient, but generated from a network with nodes randomly placed on a square surface. While much of the difference in Fig. 5.3 is superficial and due to differences in link distance, even when the links are unweighted, simulated epidemics run on these two networks show real differences (Fig. 5.4).

While the definition of clustering and its extensions to higher-order loops including four or more nodes allows us to describe important heterogeneous structures in

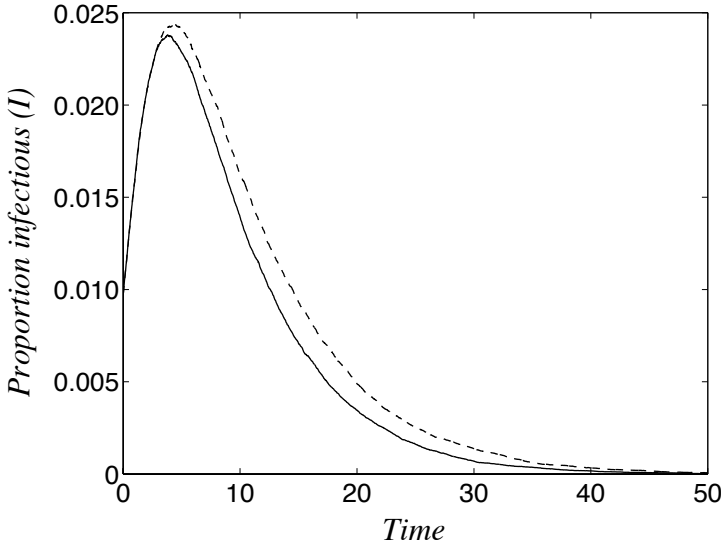


Fig. 5.4. Comparison of average of 10^4 epidemics (in the case of the Keeling clustered network, run on 100 different network realisations), on networks as illustrated in Fig. 5.3. Shown are epidemics for the Keeling clustered network (—), and for an epidemic on a triangular lattice (- - -).

networks, it does not create an analytical tool for describing the effect on disease transmission. One approach that does is *moment closure*.^{37,38} A population can be described in terms of the frequency of clusters of individuals of various types (e.g. S, I and R) and of various sizes (singlets, doublets, triplets and so on; i.e. the ‘moments’ of the distribution). By including the frequency of moments of increasingly higher order, the population can be described with increasing accuracy but at the cost of increasing complexity. Whether or not one element of a pair of susceptible individuals becomes infected, is dependent on whether one of the pair is connected to an infectious individual, i.e. if $[SS]$ is the number of $S + S$ pairs, and $[SSI]$ the number of $S + S + I$ triplets, then $\frac{d[SS]}{dt} \propto [SSI]$. Similarly $\frac{d[SSS]}{dt} \propto [SSSI]$ etc. For the simple SIR model, for example, the number of $[SI]$ pairs is determined by the equation:

$$\frac{d[SI]}{dt} = \tau[SSI] - \tau[SI] - \tau[ISI] - g[SI],$$

where $\tau[SSI]$ denotes the creation of an SI pair through the infection of S in the central position of the triplet. In a similar fashion, the number of triplets requires knowledge about the number of quadruplets, and so on. As additional accuracy is added, the system soon becomes completely intractable. However the moment closure approach offers a way of avoiding an infinite set of ordinary differential equations by ‘closing’ the system at the level of pairs and approximating triplets as

a function of pairs and individual classes.³⁷ For randomly connected networks, two different closure relations are commonly used. These differ according to the assumed error distribution under which the approximation is made. If this distribution of the error is Poisson-like, then the closure relation used is:

$$[XYZ] \approx \frac{[XY][YZ]}{[Y]}. \quad (5.4)$$

If the distribution is Bernoulli-like, then the approximation used is:

$$[XYZ] \approx \frac{\langle k \rangle - 1}{\langle k \rangle} \frac{[XY][YZ]}{[Y]}. \quad (5.5)$$

Equations (5.4) and (5.5) ignore the possible correlations between the node in state A and the node in state C , which are both in direct contact with the same node in state B . These correlations are small if the network is random. However in clustered networks there will be some heterogeneity in the probability of association between two nodes (in social networks, for example, the probability that two people will be friends will increase if they have a friend in common, or for spatially clustered populations, that the Voronoi tessellation for three nodes produces a common boundary point⁴⁰). To account for the correlation between the node in state X and the node in state Z , a modified closure relation is considered.³⁸ Let N be the total population size, and Φ the expected proportion of triplets that are triangles. Then

$$[XYZ] \approx \frac{\langle k \rangle - 1}{\langle k \rangle} \frac{[XY][YZ]}{[Y]} \left((1 - \Phi) + \frac{\Phi N}{\langle k \rangle} \frac{[XZ]}{[X][Z]} \right).$$

This approach has the attractive feature that it is transparent, easy to parameterise and builds on understanding global properties of the system based on local/neighbourhood interactions. The closure at the triplet level (i.e. ignoring loops incorporating four or more nodes) is a compromise between incorporating contact heterogeneity and retaining analytical tractability, and it has been successful in accounting for correlations that form due to diseases spreading amongst clusters of connected individuals.

In networks with even moderate levels of clustering there is a rapid decrease in the average number of new infections caused by each infectious individual. The main reason for this decline is the depletion of the susceptible neighbourhood; past the first generation, infected nodes often have at least one neighbour that is already infected. In clustered networks generated by two-dimensional spatial localisation, as described above, this is illustrated by the corresponding spatial localisation of epidemics (Fig. 5.5). While it has been shown that moment closure approximates stochastic simulations on clustered networks well,³⁸ such good agreement depends as always on the underlying model being considered. Based on a model using Poisson random networks with contact tracing and a delay before infectiousness,⁴² Fig. 5.6 shows how there is reduced agreement as clustering becomes more pronounced.

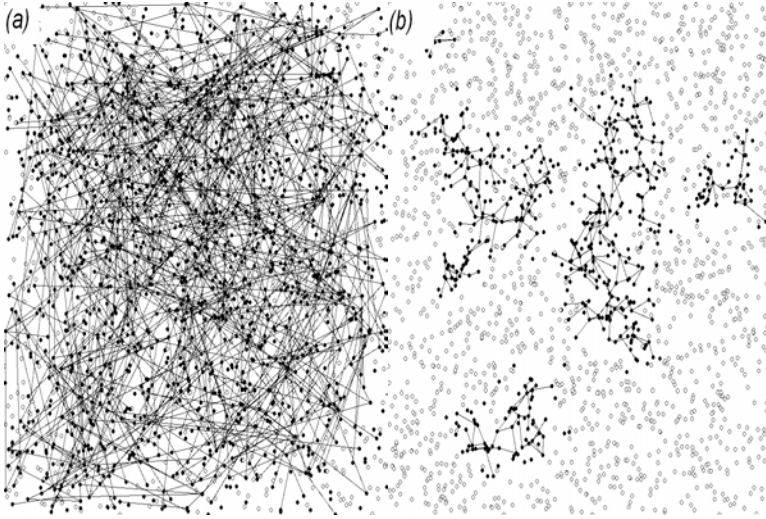


Fig. 5.5. Transmission on unclustered and spatially clustered networks. Transmission on unclustered networks fills the picture (above percolation threshold) while on clustered networks, the epidemic is self-limiting (below the percolation threshold).

While the sources of the discrepancy are not entirely clear, the delay in the onset of infectiousness and the addition of contact tracing add considerably to the complexity of the system being studied, highlighting the need for further research into analytical models of this type of contact heterogeneity. Despite these difficulties, moment closure equations as a strategic tool allow us to explore the relationship between clustering and epidemic spread,³⁸ showing how clustering can lead to a dramatic reduction in the value of R_0 if generations of infection overlap with equivalent effects on the probability of successful disease invasion. Using additional equations incorporating links between nodes along which tracing takes place, the moment closure approach can also be used to explore the effect of network dependent disease control, such as contact tracing, i.e. identifying potentially infectious connections from infected individuals.^{19,42} On a practical level, moment closure approaches have been used to explore the consequences of exploiting spatial proximity in the case of the 2001 foot-and-mouth disease epidemic,²³ as discussed in Haydon *et al.*²⁹

5.5. Networks With Heterogeneity in Contacts Per Individual

5.5.1. Models for sexually transmitted diseases

While moment closure can account for clustering, other important empirically measured network properties such as heterogeneity in contact frequency are not so easily

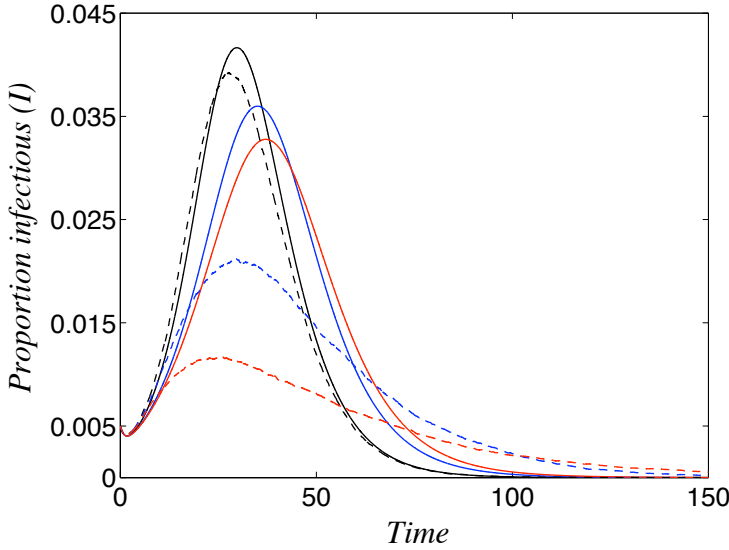


Fig. 5.6. Time evolution of the proportion of infectious nodes for moment closure equations (—) and stochastic simulations (---), for a Poisson random network with population size $N = 2000$, and $\langle k \rangle = 10$. In this simulation, infectious period is $3.5d$, latent period $3.5d$, tracing period $2d$, with a tracing rate of $2.5/\langle k \rangle$ /tracing period where d is nominally in days. Average number of infections caused by each node is $\bar{p} \times \langle k \rangle = 3.0$. Clustering coefficients are $\Phi = 0.0$ (black), 0.1 (blue) and 0.2 (red).

explored in this representation, though there are analyses that use approximations to account for them.²⁰ In sexually transmitted infections or STIs, the nature of the potentially infectious contact is well-defined, and it has long been understood that modelling their transmission and control must account for heterogeneities in sexual activity.^{5,31} Because an individual with more contacts is both more likely to be exposed to an infected individual and more likely to infect others once infected, the distribution of contacts per individual is clearly important. Assume that the probability of transmission of an STI is directly related to the number of contacts per individual, and that the population can be divided into distinct groups, with each group defined solely by the number of contacts. The number of individuals with k contacts is N_k with ($k = 1 \dots n$). For simplicity we only consider the case of a simple model in an infinite closed population. Following Ref. 5, Eqn. (5.1) can then be extended to

$$\left. \begin{aligned} \frac{dS_k}{dt} &= -\beta k S_k(t) \sum_l p(l|k) \frac{I_l(t)}{N_l} \\ \frac{dI_k}{dt} &= \beta k S_k(t) \sum_l p(l|k) \frac{I_l(t)}{N_l} - \gamma I_k(t) \end{aligned} \right\} k = 1 \dots n, \quad (5.6)$$

where S_k and I_k represent the number of susceptible and infectious individuals with k contacts, and β the per contact transmission rate between an infected and a sus-

ceptible individual. In this case frequency-dependence is used. The rate at which new infections are produced is proportional to β , the degree k of the susceptible nodes considered, the number of susceptible nodes with k connections and the probability that any given neighbour of a susceptible node with k connections is infectious. When proportionate random mixing is assumed, the probability that a node with k contacts is connected to a node with l contacts is given by $P(l|k) = lp(l) / \langle k \rangle$, where $p(l) = N_l/N$ and $\langle k \rangle = \sum_l lp(l)$ is the average number of connections in the population.

The basic reproduction number R_0 can be calculated for this system using the more general definition

$$R_0 = \lim_{N,n \rightarrow \infty} \left(\sqrt[n]{\prod_{m=1}^n I_{m+1}/I_m} \right), \quad (5.7)$$

where N is the population size, n is the generation number and I_m is the number of infected individuals in all classes in generation m .¹⁶ In this abstract model heterosexual transmission, which requires cycles of length two, is not considered. This reduces Eqn. (5.7) to: $R_0 = \lim_{N,n \rightarrow \infty} I_{n+1}/I_n$. A simple approach to calculating R_0 in this case follows.³⁶ Consider the introduction of infection into an arbitrary node in a network. This node will be of degree k with probability $p(k)$. Then for a given probability of transmission per link $\bar{p}$, the number of infected elements of an arbitrary degree l following the first generation of transmission is:

$$\begin{aligned} I_{l,1} &= \bar{p} \sum_k P(l|k) kp(k) \\ &= \frac{\bar{p}lp(l) \sum_k kp(k)}{\langle k \rangle} \\ &= \bar{p}lp(l) \end{aligned} \quad (5.8)$$

since $\langle k \rangle = \langle l \rangle$. In the following generation,

$$I_{m,2} = \bar{p} \sum_l P(m|l) I_{l,1}. \quad (5.9)$$

It is easy to show, using Eqns. (5.8) and (5.9) and summing over all node degrees, that $I_2/I_1 = I_{n+1}/I_n$ for all subsequent successive generations n and $n+1$, and therefore

$$R_0 = \bar{p} \frac{\langle k^2 \rangle}{\langle k \rangle}; \quad (5.10)$$

i.e. R_0 is proportional to the variance-to-mean ratio of the contact degree distribution in the population, where $\langle k^2 \rangle = \sum_l l^2 p(l)$ is the second moment of the contact distribution. Equation (5.10) illustrates the disproportionate role played by highly connected individuals or ‘super-spreaders’. Such models can be further

extended to account for additional properties of the population contact structure or disease characteristics, though at the cost of losing analytical tractability and model generality.

5.5.2. Disease transmission on scale-free networks

These investigations have been mirrored by equivalent investigations into social networks with high variance in degree distribution. Although random graphs have been extensively used as models of real-world networks, particularly in epidemiology, they turn out to have serious shortcomings when compared to empirical data characterising social networks such as networks of friendship within various communities, as well as networks in physical and biological systems, including food webs, neural networks and metabolic pathways. With surprising frequency, the empirically measured degree distribution is significantly different from a Poisson distribution, most importantly having a high variance-to-mean ratio. Examples include the World Wide Web, the Internet, ecological food webs, protein-protein interactions at the cellular level (e.g. Goh *et al.*²⁶), and most relevant for this discussion, human sexual networks, all with degree distributions reasonably approximated as scale-free, i.e. $p(k) \approx k^{-\gamma}$ with $2 < \gamma \leq 3$, over several orders of magnitude. As noted above, to account for the fact that each infected node past the first generation must have at least one link that ends in another infected node, the value of R_0 differs slightly from Eqn. (5.10)

$$R_0 = \bar{p}\langle k \rangle \left(\frac{\langle k^2 \rangle}{\langle k \rangle^2} - \frac{1}{\langle k \rangle} \right). \quad (5.11)$$

Note that the translation in terms of the epidemiological parameters β and γ is slightly more difficult as the depletion of links from an infected node means that the transmission rate must be increased to maintain the same R_0 ³⁹ and this in turn changes the infection rate.²⁷ While the empirically determined distribution of sexual contacts is more precisely fit with a truncated scale-free distribution,³⁴ in the limiting approximation of a scale-free infinite population with no truncation, $R_0 \rightarrow \infty$ since $\langle k^2 \rangle \rightarrow \infty$ even though $\langle k \rangle$ is finite. It follows that even an arbitrarily small transmission rate β can sustain an epidemic.⁵⁴ As implied by the name ‘scale-free’, random removal of nodes does not reduce the variance. Therefore, no amount of randomly applied, incomplete control (i.e. vaccination, quarantine) can prevent an epidemic. However, this is not the case for finite populations where the threshold behaviour is recovered⁴⁸ and targeting the small pool of highly connected nodes is sufficient to prevent an epidemic, so long as these individuals can be identified and treated or removed.

Barthélemy *et al.*⁹ showed that a further consequence of high variance distributions is the non-uniform spread of the epidemic. The higher probability that any node will be connected to a highly connected node means that disease spread follows a hierarchical order, with the highly connected nodes becoming infected first,

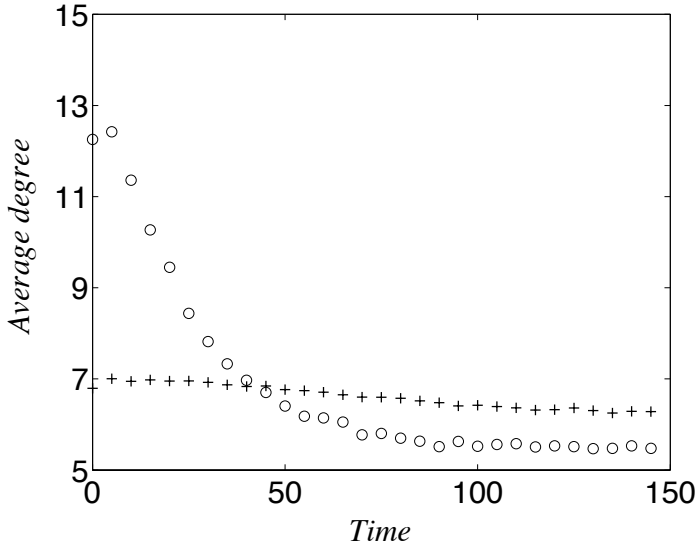


Fig. 5.7. Average degree of new infectious nodes for random (+) and truncated scale-free networks ($p(k) = Ck^{-\gamma}e^{-k/L}$ with $\gamma = 2.5$, $L = 100$ and $k \geq 3$) (o). Both networks with $N = 2000$, $\langle k \rangle = 6$. The model includes four classes (susceptible S , exposed E , infectious I , results in tracing T , and removed R) with rate of susceptibles becoming infected ($S \rightarrow E$) $0.15d^{-1}$, and, tracing occurring at rate $0.5d^{-1}$ (for all of $S \rightarrow R$, $E \rightarrow R$, $I \rightarrow R$), latent period $10d$, infectious period $3.5d$, nodes trigger tracing for $2.0d$.

and the epidemic thereafter cascading towards groups of nodes with lesser degree (Fig. 5.7 and Kiss *et al.*⁴⁴). The initial exponential growth in the time scale of epidemics is inversely proportional to the network degree fluctuations, $\langle k^2 \rangle / \langle k \rangle$. Thus the high variance in heterogeneous networks also implies an extremely small time scale for the outbreak and a very rapid spread of the epidemic, implying that in populations with these characteristics, there is a window of opportunity in epidemics when diseases can be controlled with relatively little impact on the majority of individuals (Fig. 5.8 and Kiss *et al.*⁴⁴).

However, the early infection of these nodes and the fact that they form only a small proportion of the population also means that, in a finite population, the supply of susceptible high-degree nodes is rapidly depleted. May and Lloyd⁴⁸ defined $\rho_0 = \beta \langle k \rangle / \gamma$ to be the transmission potential, equal to R_0 in homogeneously mixing (i.e. random) networks. For $\rho_0 < 1$, $R_0 < 1$ on a random network, but on a scale-free network $R_0 > 1$. For $\rho_0 > 1$, because scale-free networks lose high-degree nodes more rapidly than low-degree nodes, the variance in the degree of the remaining susceptible nodes is quickly reduced, and thus the low-degree nodes are effectively protected. Thus for sufficiently high ρ_0 , epidemics on random networks last longer, and also are able to reach more nodes. Above a certain value ρ_{crit} , the final epidemic

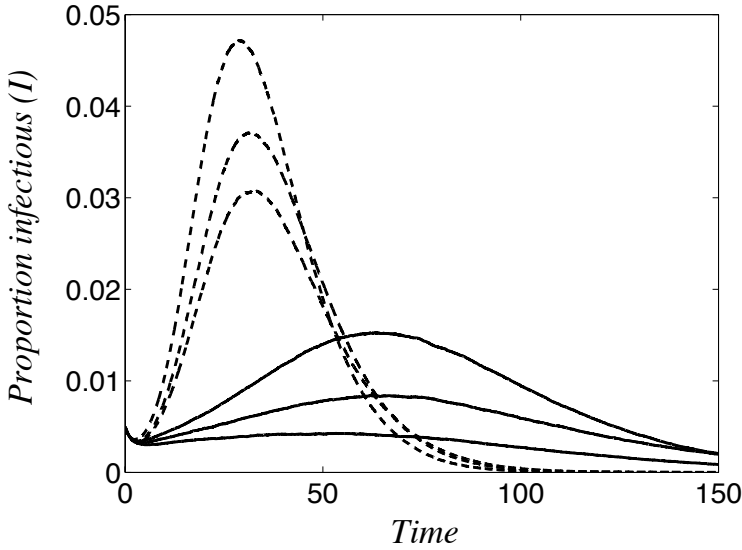


Fig. 5.8. Time evolution of the proportion of infectious nodes for random (—) and truncated scale-free networks ($p(k) = Ck^{-\gamma}e^{-k/L}$ with $\gamma = 2.5$, $L = 100$ and $k \geq 3$) (---), where $N = 2000$, $\langle k \rangle = 6$, for epidemics with infection rates per link $\beta = 0.067, 0.0735, 0.08$. Latent period is $3.5d$, infectious period $3.5d$.

size on random networks is larger^{43,48} and as $\rho_0 \rightarrow \infty$, approaches its asymptote (the total population size) more rapidly than for scale-free networks (Fig. 5.9).

5.5.3. Preferential attachment or the ‘Matthew effect’

The common appearance of scale-free structures in both nature and human endeavour is suggestive that universal laws are in operation, which, if understood, could be exploited in controlling disease. Networks mimicking scale-free type degree distributions can be generated using the preferential attachment model proposed by Barabási and Albert⁸ (or BA model) as a possible reason behind many of these structures. In social science, this is sometimes known as the ‘Matthew effect’[‡] which can effectively be described as ‘the rich get richer’. The network construction algorithm starts with a small number (m_0) of connected nodes. At every step, a new node with $m(\leq m_0)$ links is added to the network, connecting to already existing nodes. The probability Π that a new node connects to an existing node u depends on the degree of that node with $\Pi(u_k) = u_k / \sum_l u_l$. Numerical simulations of the Barabási and Albert model produce networks that well approximate a scale-free degree distribution with exponent $\gamma = 2.9 \pm 0.1$. The analytical expression for the

[‡]For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.’ (Matthew XXV:29, King James Bible.)

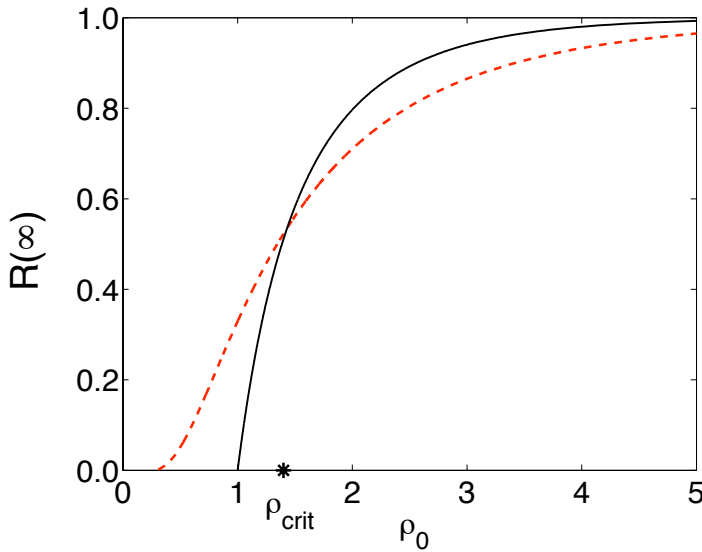


Fig. 5.9. Final epidemic size $R(\infty)$ as a function of the transmission potential ρ_0 computed analytically for the mean-field SIR model (—) and semi-analytically for Barabási-Albert or BA networks (- - -). For the BA networks $R(\infty)$ increases from close to zero, however for the mean-field case it only increases from $\rho_0 = 1$. The value of $R(\infty)$ for the scale-free network increases more slowly, however, due to the depletion of highly connected nodes.

degree distribution $p(k) = \frac{2m_0^2}{k^3}$ gives a value of $\gamma = 3$, independent of the original starting value m_0 . While preferential attachment is unlikely to directly explain the distribution in sexual contact networks, for example, it is certainly possible that experience gained from successfully establishing contacts can improve the probability of success, thus mimicking the preferential attachment mechanism to some degree.

5.5.4. STI partnership models

In the simplest network models the connections of the population are fixed with no switching of links; in contrast, Kermack–McKendrick type models can be viewed as populations where the links switch at an infinitely rapid rate.⁵³ Of interest is the interaction between the two extremes, i.e. when the dynamics of the network changes the dynamics of disease. While we shall not deal with this theme extensively, the concurrency of links has received considerable study^{18,20,25,52,61} in the modelling of STIs, where the nature of the partnerships between individuals is emphasised, rather than the individuals themselves. This dyad-based approach often assumes that epidemic dynamics are driven by serially monogamous relationships.^{18,52} Despite this abstraction, they are of interest because of the emphasis on the dynamics of the network itself – in the simplest case, no epidemic can occur if all partnerships are sufficiently long. The networks generated from partnership models illustrate the

importance of both ‘traditional’ static network properties, for example number of partners and network structures such as the centrality of an individual in a network, as well as dynamic properties such as the concurrency of partnerships.

Whether an individual’s likelihood of becoming infected, or if infected, his likelihood of being important for transmission has been shown to depend differently on network properties, at least for some systems believed to be relevant for STIs.²⁵ In the first case, the number of individuals by whom that individual could be infected is most important (i.e. the in-degree of the individual); in the second case, the ‘depth’ of network paths from that individual, as determined by the path length distribution and global measures, such as node centrality (e.g. betweenness, which is a measure of how often an individual is part of the most efficient path connecting other individuals in a network).

5.6. Integrating Networks and Epidemiology

Thus far we have considered the properties of the social network of potentially infectious contacts, i.e. which nodes a node could infect, if it were infectious. This is important and often the only logical approach if, for instance, no disease data are available or if the properties of the underlying social network are being exploited for disease control. For example, for the purposes of analysing the efficacy of tracing potentially infectious contacts for disease control, the social network can be vital.^{19,32,42} However, in the absence of control or when control is not based on exploiting social network structure, given a contact network and the characteristics of a disease that can spread on the network, one can thin links to generate the network of truly infectious links (as disease will not necessarily spread across all available links), referred to as the *transmission* or *epidemiological* network. Such a network is inherently directed (since one must consider separately the probability of infection in each direction) even when the social network is undirected, however, the thinned network is usually significantly more sparse. Further, while the social network may have weightings attached to links and nodes, the epidemiological network is unweighted so long as the infectious state of any node is not dependent on any network parameters (e.g. one cannot have a node that is more infectious if it has been infected by exposure to multiple infected neighbours).

It is also often the case that networks generated with different disease assumptions will have different properties from the underlying social network. For example, following Trapman,⁵⁹ consider two systems in which both have a constant infectiousness per link per unit time $\tau(t)$ but with either fixed infectious periods θ_A (system A) or bimodal infectious periods, with a proportion $1 - X$ with a zero infectious period and proportion X with an infectious period of length θ_B (system B), such

that

$$\bar{p}_{av} = \int_0^{\theta_A} \tau(t) dt = X \int_0^{\theta_B} \tau(t) dt, \quad (5.12)$$

i.e. for the two systems the average probability of infection per link $\bar{p}_{av}$ is the same. This latter system B can be thought of as a population where only some individuals are susceptible to disease. In system A , there is a fixed probability of transmission per link – in this case, the epidemic threshold $R_0 = 1$ corresponds to the *bond percolation* threshold (i.e. all sites occupied, but links present only with the probability $\bar{p}_{av}$). In system B , consider the limit where $\theta_B \rightarrow \infty$. Then the individuals in the proportion X are able to transmit with 100% probability, while the remainder never do. As $\bar{p}_{av}$ increases, X increases and $R_0 = 1$ corresponds to the *site percolation* threshold. Similarly, perfect vaccination could be viewed as having an effect on the site percolation of the *original* epidemiological network, removing whole nodes from the network, and thus the most relevant question is the coverage required, i.e. how many individuals must be vaccinated. Imperfect vaccination however, is more related to bond percolation, if it is assumed there is perfect coverage but imperfect protection.

5.6.1. Component sizes and the final epidemic size

In a network, disease may continue to spread so long as an infected node can reach at least one uninfected node. A component represents a subset of nodes in which all nodes can reach each other. The largest such component is called the giant component. In many real-world networks, edges/links are directed, for example the Internet, the World Wide Web (e.g. webpage B can be accessed via hyperlinks from webpage A with the reciprocal not being true), or where movement of individuals carries the disease (e.g. one-way movements of individuals between cities, or of livestock between farms). Therefore two components are now of interest: the strongly connected components or strong components represented by subsets of the directed network in which all nodes can reach each other in both directions, and weakly connected components or weak components which are strong components plus all its sources and sinks.^{51,58} In an epidemiological network, any disease starting in a strong component or at a source node will infect all elements of the strong component and all sink nodes, but not necessarily all sources. Thus, the largest or giant strongly connected component (GSCC), in the absence of any interventions or control measures, is an estimate of the lower bound of the maximum epidemic size, while the giant weakly connected component is an estimate of its upper bound (e.g. Ref. 35).

5.6.2. R_0 on epidemiological networks and network percolation thresholds

The epidemiological network allows us to establish a connection between the network percolation threshold and R_0 . In a randomly mixed epidemiological network, R_0 is the network percolation threshold,^{12,58} loosely defined as the point at which the final epidemic size is expected to scale with the size of the population (discussed in Ref. 35).

The result of Eqn. (5.10) can be easily extended to consider weighted directed links and with variable susceptibility of nodes it can also be shown that

$$R_0 = \bar{p} \frac{\langle \tau k_{out} \sigma k_{in} w \rangle}{\langle \tau k_{out} w \rangle} \quad (5.13)$$

where τ and σ are the weighting of the out- and in-links, w the weighting associated with each node, k_{in} the number of inward links and k_{out} the number of outward links.^{35,58} Note that in Eqn. (5.11), the node at the end of one of the links after the initial generation is already infected, while in Eqn. (5.13), this does not occur because the in-links and out-links are distinct. In this case, the equation for R_0 reduces to $R_0 = \frac{\langle l_{in} l_{out} \rangle}{\langle l_{out} \rangle}$ in the epidemiological network generated from a directed network where nodes have uncorrelated in- and out-links or a network with dynamic links, or $R_0 = \frac{\langle l_{in} l_{out} \rangle}{\langle l_{out} \rangle} - \frac{\bar{p}^2}{\langle l_{out} \rangle}$ when generated from static networks, where l_{in} and l_{out} are the number of inward and outward ‘truly infectious’ links per node and $\bar{p}^2$ arises as the probability that an undirected potentially infectious link generates transmission links in both directions.

While this approach is only valid for randomly connected networks, it can be useful in other contexts, provided a network can be transformed into a randomly-connected structure. We illustrate this in the case of the small-world network for which both the bond and site percolation threshold problems have been solved.⁵⁰ In the absence of long-range connections, increases in the transmission probability per link will result in the growth of local clusters in the epidemiological network that would correspond to the local epidemic size, should an element in that cluster become infected (Fig. 5.10). In the simplest case of a one-dimensional small-world lattice (i.e. with all nodes having local connections to exactly two neighbours), the probability $\hat{p}_C$ that a local cluster of infected individuals will be of size C depends in a straightforward fashion on the probability $\bar{p}$ that a given link is infectious, if one assumes that, during the initial spread of the disease, the probability of a long-range link returning to an already infected cluster is small. Then in this case, $\hat{p}_C = (1 - \bar{p})^2 \bar{p}^{C-1}$ since the two end links must be non-infectious and all other $C - 1$ links in the cluster must be infectious. Moore and Newman⁵⁰ use the expression for the local cluster size to determine the percolation threshold via a direct calculation based on the number and size of clusters connected by long-range shortcuts. Another approach is to construct an epidemiological network (with directed links) and contract all nodes in a local cluster into a single ‘supernode’.

The probability that there will be a supernode of size C in the (now directed) epidemiological network is $p_C = C(1 - \bar{p})^2 \bar{p}^{C-1}$; e.g. for a cluster of size $C = 3$, with three consecutive nodes X, Y and Z , one could have a cluster of size C with $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$ or $X \leftarrow Y \leftarrow Z$. Each supernode will have an average of $\bar{p}qC$ infectious long-range connections if the probability of a node having a long-range connection in the original network was q . For a sufficiently large population, with all clusters contracted into supernodes, the resultant network of supernodes is randomly connected, and so Eqn. (5.13), while not equal to R_0 , is the epidemic percolation threshold of the network. Therefore what one might call R_0^{SN} (i.e. for the system of supernodes) reduces to

$$\begin{aligned} R_0^{SN} &= \bar{p}q \sum_{C=1}^{\infty} C p_C \\ &= (1 - \bar{p})^2 q \sum_{C=1}^{\infty} C^2 \bar{p}^C \\ &= q\bar{p} \frac{(1 + \bar{p})}{(1 - \bar{p})}. \end{aligned} \tag{5.14}$$

The expression for the distribution of local cluster sizes becomes significantly more complicated for higher-dimensional small-world networks, however the principle remains the same. The interpretation of local clusters linked by long-range connections is closely related to a household model of disease transmission, in which the distribution of epidemic sizes within households is used to generate the value of the between-household value of R_0 . Figure 5.10 shows the epidemiological network corresponding to the small world network of Fig. 5.2 where 50% of links are considered infectious – in this case, development of the linked clusters can clearly be seen.

5.6.3. *Contact frequency distributions on social and epidemiological networks*

Epidemiological network structure can differ considerably from the social network structure due to link weightings. Following an idea developed in Ref. 36, consider a network of individuals linked by sexual contacts. In an illustrative toy model of an STI, we account for heterogeneity (i.e. high variance in the number of contacts) by using the BA scale-free network model as previously described. We assume that the network is static.

The number of sexual partners and duration of partnership are often inversely correlated.^{25,52} To reflect this, we assign a weighting to each link by assuming that the probability that the strength of interaction through a sexual partnership between two individuals is inversely proportional to the number of partners of the individuals, i.e. $\frac{1}{\text{Degree}(A) * \text{Degree}(B)}$ and that the probability of transmission of an STI is directly proportional to this quantity. We then use this relationship to

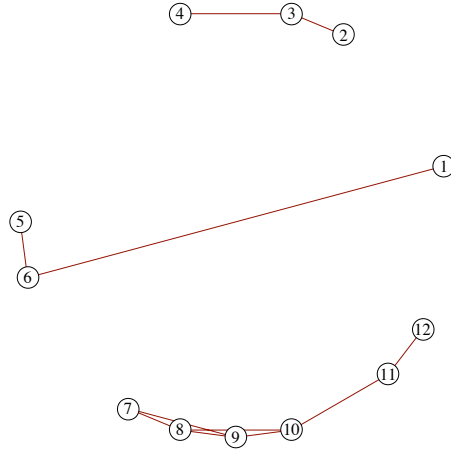


Fig. 5.10. Epidemiological network generated from the small-world model, with 50% of links considered infectious. Clusters are formed by nodes as (1), (2,3,4), (5,6), (7,8,9,10), and (11,12) with long-range infectious links joining nodes 1 to 6 and 10 to 11.

build epidemiological networks. Depending on the type of disease or transmission mechanism, per contact probability of transmission can be different. To illustrate this we construct epidemiological networks such that only links with a probability greater than a set threshold are accepted, i.e. $\frac{1}{\text{Degree}(A) * \text{Degree}(B)} > p_{th}$. In each epidemiological network the degree distribution is illustrated in Fig. 5.11. The expected degree distribution in the epidemiological network is then

$$q(m) = \sum_k p(k) \Omega \left(m, \left(\frac{jp(j)}{\langle k \rangle} z (1 - z) \right)_j \right), \quad (5.15)$$

$$z = \frac{1}{jkA}. \quad (5.16)$$

Here $q(m)$ represents the degree distribution in the epidemiological network and run over all degrees in the social network. The distribution Ω denotes the proportion of successful trials obtained from events occurring with probability $\frac{jp(j)}{\langle k \rangle}$ and an associated probability of success $\frac{1}{jk}$. The probability is normalised by $A = \sum_{E(j,k)} \frac{1}{jk}$, where the weights are summed over all edges in the social network.

For the same underlying contact network, depending on the transmission threshold (i.e. a surrogate for different disease types or transmission mechanism), the epidemiological network has very different properties. The most striking effect is the limited role played by highly connected nodes in the transmission process. There

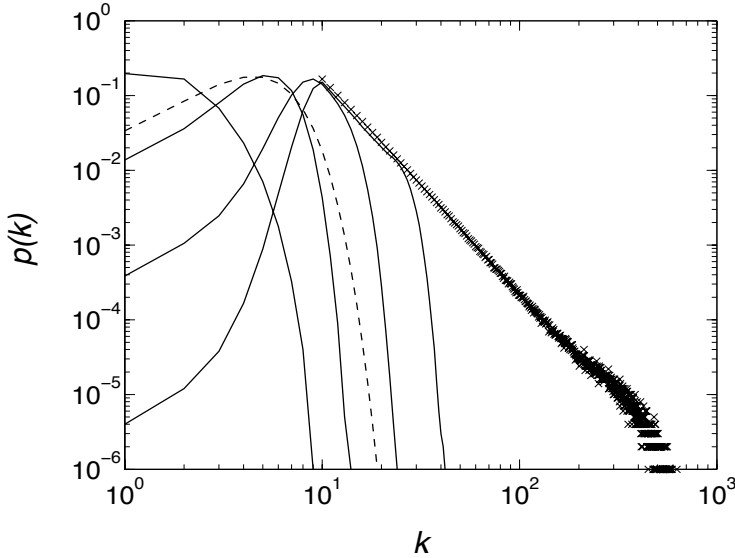


Fig. 5.11. The degree distribution of the epidemiological networks generated from a BA scale-free social network for link weightings in the social network that are inversely proportional to the degrees of the nodes connected. As the probability of transmission decreases, the variance in the infected nodes decreases. For comparison, the degree distribution of a random network is shown (dashed line).

are also considerable differences between the different epidemiological networks, conceptually illustrating different types of diseases or different transmission mechanisms. This is highlighted by plotting R_0 (Fig. 5.12) as defined by Eqn. (5.13) and with the distribution defined by Eqn. (5.15) for the different epidemiological networks, while recalling that, for a true scale-free network, R_0 is infinite for any fixed infectiousness per link greater than zero. These estimates are approximate, as the $1/(kl)$ weighting introduces strong correlations between nodes that are poorly connected and thus the network is no longer randomly connected, so Eqn. (5.13) might not be entirely appropriate. However, the relationship between the measured social network degree distribution and epidemiological weightings, resulting in much lower variance (and thus R_0), highlights the importance of understanding the epidemiological question when examining the social structure. In the case of HIV, for example, the effect of multiple exposures in long-term partnerships is mitigated by the relatively short infectious period. The number of partnerships, not the number of acts, remains the key epidemiological parameter.⁴ There is recent evidence, however, that the virus strain HIV-1 may be evolving towards lower viral replicative fitness,⁶ suggesting decreased pathogenicity of HIV-1 over time. However, if lower pathogenicity (presumably resulting in a lower probability of transmission per act) is accompanied by a longer infectious period, individuals involved in relatively few

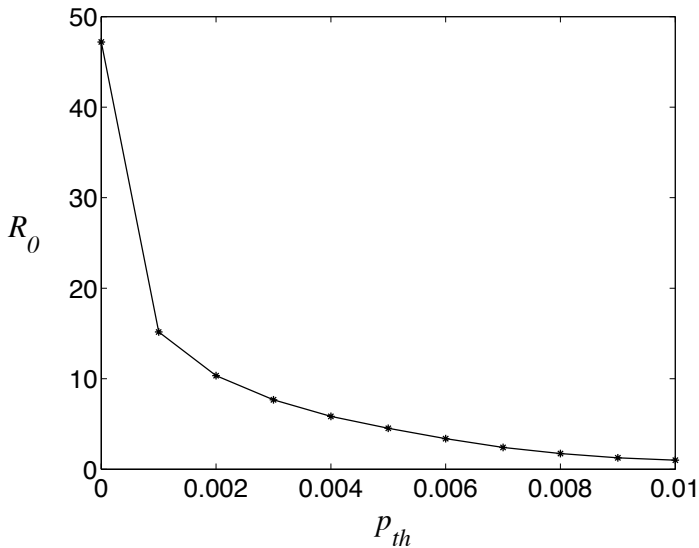


Fig. 5.12. Calculated values of R_0 for epidemiological networks, showing the dramatic decrease in R_0 as the transmission probability p_{th} increases. Link strength is inversely weighted to the degrees of the connected nodes.

longer-term partnerships with greater exposure would have an increased risk of infection per partnership than individuals involved in many short-term partnerships. This would result in epidemiological networks where highly connected individuals have a less important role than individuals involved in fewer partnerships but with more sexual interactions across these contacts (as in Fig. 5.11). Thus while the social network pattern is unchanged, changes in the transmission characteristics may result in a different epidemiological network involving potential shifts in risk, and therefore in the focus of control strategies.

5.7. Conclusion

In this chapter, we have illustrated a few simple points regarding the interplay between two rich subject areas, disease dynamics and social network analysis. While the history of mathematical epidemiology contains many of the ideas that have since been replicated in social network theory, the study of social networks has generated both new ideas and new impetus to understanding the role that contact heterogeneity can play in the spread, persistence and control of infectious diseases. We offer our apologies to the authors of many valuable and interesting papers originating from both traditions that we have omitted; however, rather than presenting an exhaustive study of the results from either, we have concentrated instead on presenting illustrations of how disease dynamics can only be properly understood

by considering a combination of both pattern and process. Critical to this is the interplay of individuals from both traditions, who will bring together the analytical strengths and insights they both have to offer (e.g. Ref. 10).

References

1. R. Albert, H. Jeong, and A.-L. Barabási, Diameter of the World-Wide web, *Nature*. **401**, 130 – 131, (1999).
2. R. Albert, H. Jeong, and A.-L. Barabási, Error and attack tolerance of complex networks, *Nature*. **406**, 308 – 382, (2000).
3. R. Albert, and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74**, 47 – 97, (2002).
4. R.M. Anderson, and R.M. May, Epidemiological parameters of HIV transmission, *Nature*. **333**, 514 – 9, (1988).
5. R.M. Anderson, and R.M. May, *Infectious Diseases of Humans: Dynamics and Control*. (Oxford University Press, 1992).
6. K.K. Arien, R.M. Troyer, Y. Gali, R.L. Colebunders, E.J. Arts, and G. Vanham, Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time, *Aids*. **19**, 1555 – 64, (2005).
7. F. Ball, D. Mollison, and G. Scalia-Tomba, Epidemics with two levels of mixing, *Annals of Applied Probability*. **7**, 46 – 89 (1997).
8. A.-L. Barabási, R. Albert, Emergence of scaling in random networks. *Science*. **286**, 509 – 12 (1999).
9. M. Barthelemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, Velocity and hierarchical spread of epidemic outbreaks in scale-free networks, *Phys. Rev. Lett.* **92**, 178701 (2004).
10. S. Bansal, B.T. Grenfell, and L.A. Meyers, When individual behaviour matters: homogeneous and network models in epidemiology, *J. Roy. Soc. Interface*. **4**, 879 – 891, (2007).
11. B. Bolker, and B.T. Grenfell, Space, persistence and dynamics of measles epidemics, *Philos Trans R Soc Lond B Biol Sci*. **348**, 309 – 20, (1995).
12. R. Cohen, D. Ben-Avraham, and S. Havlin, Percolation critical exponents in scale-free networks, *Phys Rev E*. **66** (3 Pt 2A):036113, (2002).
13. F. Chung, and L. Lu, The diameter of sparse random graphs, *Adv. Appl. Math.* **26**, (2001).
14. L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, Comparing community structure identification, *J. of Stat. Mech.* **P09008**, (2005).
15. O. Diekmann, and J.A.P. Heesterbeek, *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. (Mathematical and Computational Biology. New York: John Wiley & Sons, 2000).
16. O. Diekmann, J.A.P. Heesterbeek, and J.A.J. Metz, On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365 – 382, (1990).
17. R. Durrett, and S.A. Levin, The importance of being discrete (and spatial), *Theor. Popul. Biol.* **46**, 363 – 394, (1994).
18. K. Dietz, and K.P. Haderler, Epidemiological models for sexually transmitted diseases, *J. Math. Biol.* **26**, 1 – 25, (1998).
19. K.T. Eames, and M.J. Keeling, Contact tracing and disease control, *Proc. Roy. Soc. B*. **270**, 2565 – 71, (2003).

20. K.T. Eames, and M.J. Keeling, Monogamous networks and the spread of sexually transmitted diseases, *Math. Biosci.* **189**, 115 – 30, (2004).
21. P. Erdős, and A. Rényi, On Random Graphs, *Publ. Math. Debrecen.* **6**, 290 – 297, (1959).
22. S. Eubank, H. Guclu, V.S. Kumar, M.V. Marathe, A. Srinivasan, Z. Toroczka, and N. Wang, Modelling disease outbreaks in realistic urban social networks, *Nature.* **429**, 180 – 4, (2004).
23. N.M. Ferguson, C.A. Donnelly, and R.M. Anderson, The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions, *Science.* **292**, 1155 – 1160, (2001).
24. N.M. Ferguson, D.A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D.S. Burke, Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature.* **437**, 209 – 14, (2005).
25. A.C. Ghani, J. Swinton, and G.P. Garnett, The role of sexual partnership networks in the epidemiology of gonorrhea, *Sex. Transm. Dis.* **24**, 45 – 56, (1997).
26. K.I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, Classification of scale-free networks, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12583 – 8, (2002).
27. D.M. Green, I.Z. Kiss, and R.R. Kao, Parameterisation of Individual-Based Models. *J. Theor. Biol.* **236**, 289 – 297, (2006).
28. B.T. Grenfell, O.N. Bjornstad, and J. Kappey, Travelling waves and spatial hierarchies in measles epidemics, *Nature.* **414**, 716 – 723, (2001).
29. D.T. Haydon, R.R. Kao, and P. Kitching, On the aftermath of the UK Foot-and-Mouth Disease outbreak, *Nature Reviews Microbiology.* **2**, 675 – 681, (2004).
30. J.A.P. Heesterbeek, and M.G. Roberts, The type-reproduction number T in models for infectious disease control, *Math. Biosci.* **206**, 3 – 10, (2007).
31. H.W. Hethcote, J.A. Yorke, and A. Nold, Gonorrhea modeling: a comparison of control methods, *Math. Biosci.* **58**, 93 – 109, (1982).
32. R. Huerta, and L.S. Tsimring, Contact tracing and epidemics control in social networks, *Phys. Rev. E.* **66**, 056115, (2002).
33. H.J. Jones, and M.S. Handcock, An assessment of preferential attachment as a mechanism for human sexual network formation, *Proc. R. Soc. Lond. B.* **270**, 1123 – 1128, (2003).
34. J.H. Jones, and M.S. Handcock, Social networks: Sexual contacts and epidemic thresholds, *Nature.* **423**, 605 – 6, (2003).
35. R.R. Kao, L. Danon, D.M. Green, and I.Z. Kiss, Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain, *Proc. R. Soc. B.* **273**, 1999 – 2007, (2006).
36. R.R. Kao, Evolution of Pathogens towards low R_0 . *J. Theor. Biol.* **242**, 634 – 642 (2006).
37. M.J. Keeling, D.A. Rand, and A.J. Morris, Correlation models for childhood epidemics, *Proc. R. Soc. B.* **264**, 1149 – 1156, (1997).
38. M.J. Keeling, The effects of local spatial structure on epidemiological invasions, *Proc. R. Soc. B.* **266**, 859 – 67, (1999).
39. M.J. Keeling, and B.T. Grenfell, Individual-based perspectives on R_0 , *J. Theor. Biol.* **203**, 51 – 61, (2000).
40. M.J. Keeling, M.E.J. Woolhouse, D.J. Shaw, L. Matthews, M. Chase-Topping, D.T. Haydon, S.J. Cornell, J. Kappey, J. Wilesmith, and B.T. Grenfell, Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape, *Science.* **294**, 813 – 817, (2001).

41. W.O. Kermack, and A.G. McKendrick, A contribution to the mathematical study of epidemics, *Proc. R. Soc. London Ser. A.* **115**, 700 – 721, (1927).
42. I.Z. Kiss, D.M. Green, and R.R. Kao, Disease contact tracing in random and clustered networks, *Proc. R. Soc. B.* **272**, 1407 – 14, (2005).
43. I.Z. Kiss, D.M. Green, and R.R. Kao, The effect of contact heterogeneity and multiple routes of transmission on final epidemic size, *Math. Biosci.* **203**, 124 – 36, (2006).
44. I.Z. Kiss, D.M. Green, and R.R. Kao, Disease Contact Tracing in Random and Scale-Free Networks, *J. Roy. Soc. Interface.* **3**, 55 – 62, (2006).
45. S.A. Levin, and R. Durrett, From individuals to epidemics, *Phil. Trans R. Soc. London B.* **351**, 1615 – 1621, (1996).
46. R. Levins, Some demographic and genetic consequences of environmental heterogeneity for biological control, *Bull. Entomol. Soc. Am.* **15**, 237 – 240, (1969).
47. F. Liljeros, C.R. Edling, L.A. Amaral, H.E. Stanley, and Y. Aberg, The web of human sexual contacts, *Nature.* **411**, 907 – 908, (2001).
48. R.M. May, and A.L. Lloyd, Infection dynamics on scale-free networks, *Phys. Rev. E.* **64**, 066112, (2001).
49. L.A. Meyers, M.E.J Newman, M. Martin, and S. Schrag, Applying Network Theory to Epidemics: Control Measures for Mycoplasma pneumoniae Outbreaks, *Emerging Infectious Diseases.* **9**, 204 – 210, (2003).
50. C. Moore, and M.E.J Newman, Exact solution of site and bond percolation on small-world networks, *Phys. Rev. E.* **62**, 7059-64, (2000).
51. M.E.J Newman, S.H. Strogatz, and D.J. Watts, Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E.* **64**, 026118, (2001).
52. M. Morris, and M. Kretzschmar, Concurrent partnerships and the spread of HIV, *Aids.* **11**, 641 – 8, (1997).
53. P.E. Parham, and N.M. Ferguson, Space and contact networks: capturing the locality of disease transmission, *J. R. Soc. Interface.* **3**, 483 – 93, (2006).
54. R. Pastor-Satorras, and A. Vespignani, Epidemic spreading in scale-free networks, *Phys. Rev. Lett.* **86**, 3200, (2001).
55. M. Roberts, and H. Heesterbeek, Bluff your way in epidemic models, *Trends Microbiol.* **1**, 343 – 348, (1993).
56. M.G. Roberts, and J.A.P. Heesterbeek, A new method for estimating the effort required to control an infectious disease, *Proc. Biol Sci.* **270**, 1359 – 1364, (2003).
57. R. Ross, *The Prevention of Malaria*, (2nd edn., Churchill, London, 1911).
58. N. Schwartz, R. Cohen, D. ben-Avraham, A.-L. Barabási, and S. Havlin, Percolation in directed scale-free networks, *Phys. Rev. E.* **66**, 015104(R), (2002).
59. P. Trapman, On analytical approaches to epidemics on networks, *Theor. Popul. Biol.* **71**, 160 – 173, (2007).
60. P. van den Driessche, and J. Watmough, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, *Math. Biosci.* **180**, 29 – 48, (2002).
61. C.H. Watts, and R.M. May, The influence of concurrent partnerships on the dynamics of HIV/AIDS, *Math. Biosci.* **108**, 89 – 104, (1992).
62. D.J. Watts, and S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature.* **393**, 440 – 442, (1998).

Chapter 6

Evolutionary Origin and Consequences of Design Properties of Metabolic Networks

THOMAS PFEIFFER¹ AND SEBASTIAN BONHOEFFER²

¹*Program for Evolutionary Dynamics, Harvard University*

²*Institute of Integrative Biology, ETH Zurich*

pfeiffer@fas.harvard.edu, sebastian.bonhoeffer@env.ethz.ch

Processes in living systems are the result of interacting biochemical compounds in highly complex biochemical reaction networks. Genomic data allow reconstruction of these networks and analysis of their design properties. It is a major challenge in biology to understand the origin and consequences of these design properties. Since biochemical reaction networks are the result of evolution, it is a promising approach to study the impact of evolutionary processes on network design. Conversely, network design may influence network evolution, because it determines the relation between genotype, environment and phenotype of an organism. Here we describe approaches to studying the evolutionary origin and consequences of key properties of metabolic networks.

6.1. Introduction

As one of the best-studied network types in biology, analysing metabolism in the context of evolution has considerable advantages compared to other biochemical networks such as signal transduction or gene regulation networks.

Firstly, there is a large body of experimental data on metabolism. For most biochemical reactions, the corresponding enzyme is known and sequence data are available (see, for example, www.genome.ad.jp/kegg¹). On the basis of theoretical methods such as Flux Balance Analysis (FBA) and Elementary Modes Analysis,² these data allow reconstruction of many properties of metabolic networks, particularly of organisms with completely sequenced genomes.^{3–5} High-throughput techniques can be used to quantify properties of metabolic networks, such as enzyme expression patterns, flux distributions or metabolite concentrations.^{6–10} Additionally, in a number of well-studied metabolic subsystems, for example amino acid synthesis, glycolysis and oxidative phosphorylation, kinetic properties of the involved enzymes are known (see, for example, www.brenda.uni-koeln.de¹¹). The detailed knowledge on metabolism provides an excellent basis for relating the phenotypic

properties of an organism to its genotype.

Secondly, there are well-developed theoretical methods to define, describe and analyse properties of metabolism (see, for example, Ref. 12). These methods are based on two different approaches, often referred to as the *stoichiometric* and the *kinetic* approaches.¹³ The stoichiometric approach is used to analyse topological properties of metabolic networks based on stoichiometry, i.e., the information of how metabolites are transformed into each other by biochemical reactions. The main advantage of the stoichiometric approach (and simultaneously its major limitation) is that no knowledge about kinetic properties of the biochemical reactions is required. Therefore it can be applied to large metabolic reaction networks, where all biochemical reactions but not all relevant kinetic data are known. Consequently, stoichiometric approaches such as Elementary Modes Analysis and FBA are essential in the reconstruction of metabolic networks from genomic data.^{2–5} On the other hand, kinetic approaches such as Metabolic Control Analysis (MCA) play an important role in incorporating and analysing kinetic features of metabolic systems.^{12,14} The kinetic approach is essential for quantitative descriptions and predictions of the temporal dynamics of metabolic networks. Applied in an evolutionary context, both types of theoretical approaches can help to explain patterns observed in metabolic systems and to derive predictions for their evolution.

Thirdly, the evolution of key properties of metabolism can be directly observed in experimental evolution studies on microbial populations. The relative simplicity of microbes such as yeast and *E. coli* allows manipulation of metabolic properties and determination of the relationship between metabolic properties and fitness.¹⁵ Their small size and fast reproduction cycle allows evolutionary changes to be observed in large populations for thousands of generations (see, for example, Ref. 16). In the context of metabolism, a number of long-term evolution studies resulted in interesting and unexpected observations. Long-term evolution experiments on *E. coli* in continuous culture (chemostat), for example, show that stable polymorphisms may evolve in microbial populations that are limited by a single resource. These polymorphisms are not expected on the basis of the competitive exclusion principle. It could be shown that they were maintained by crossfeeding interactions, where one strain degrades the limiting substrate only partially and excretes a product that can be used as a substrate by a second strain.^{17–19} Long-term evolution experiments in batch culture indicate that populations adapt towards optimal flux distribution patterns as predicted by FBA.²⁰ Interestingly, the rate of adaptation was faster in organisms that had previously been disturbed by knockout mutations. Finally, the high flexibility of microbial metabolism that often allows usage of a large range of different substrates results in a high diversity of metabolic properties that can be selected in an appropriate environment, and the existence of alternative metabolic pathways with the same biochemical function allows studies on the advantages of specific properties of an alternative pathway in a given environment.¹⁷

In summary, metabolism is an ideal system for studying evolutionary phenomena

and, conversely, evolutionary biology may offer valuable approaches to studying metabolic systems. In the following we discuss theoretical approaches to studying the evolution of metabolism. We first review theoretical studies on optimal design of metabolic systems. In these studies, simplified models of metabolic pathways are used to analyse key properties such as optimal enzyme expression or optimal reaction orders. Furthermore, they allow conclusions to be derived on properties of metabolic systems that are of relevance to their evolution, such as robustness and epistasis. Finally, we present novel approaches to studying the evolutionary origin of large-scale design properties in metabolic networks and their evolutionary consequences.

6.2. Optimal Design of Metabolic Pathways

Studies that focus on the question of how evolution affects kinetic properties of existing pathways often apply optimisation principles to the design of metabolic pathways. The following kinetic properties of metabolic pathways are considered as being under selection pressure: (i) the flux through the pathway is maximised, (ii) yield is maximised, (iii) enzyme concentrations are minimised, (iv) intermediate concentrations are minimised. Often, these properties depend on each other and cannot be optimised simultaneously. Evidence that the above properties are of importance in the evolution of metabolic pathways has been discussed by Heinrich and Schuster.¹²

A simple but revealing approach to derive optimal properties of ATP-producing pathways has been proposed by Waddell and co-workers.²¹ Using linear flux-force relation to describe the dependence of the flux of a pathway on the free energy difference between substrates and products, it can be shown that the energy yield that maximises the rate of ATP production is 0.5, i.e. half of the free energy difference between substrate and product is conserved as ATP and half is used to drive the pathway. With increasing energy yield, the rate of ATP production decreases and thus a trade-off exists between rate and yield of ATP production. However, the applicability of a linear flux-force relation to biochemical pathways has been questioned, as it is often not compatible with common kinetic descriptions of biochemical reactions.¹² On the other hand, theoretical studies that are based on an explicit kinetic description of the mechanisms of ATP production result in similar findings for the optimal design of glycolysis and thus support the above approach.²² Additionally, these studies allowed the prediction of the optimal order of reactions in ATP-producing pathways. In line with observed patterns in glycolysis it has been predicted that, against common intuition, ATP-consuming reactions in the upper part of an ATP-producing pathway may increase the rate of ATP production. ATP-producing reactions are correctly predicted to be located in the lower part of the pathway. Thus it seems to be advantageous to *invest* energy into the beginning of a pathway.

An analogous finding is obtained when maximising the rate of a pathway (not necessarily an ATP-producing pathway) under constraints for the total concentration of enzymes.¹² Here, it has been obtained that a larger amount of enzyme should be allocated into the reactions in the upper part of a pathway compared to the reactions in the lower part. For a linear pathway of enzymes with irreversible kinetics, it has in fact been derived that the maximally possible amount of enzyme should be allocated into the first reaction, as the rate of an irreversible pathway is completely determined by the first step. However, in this case, intermediate concentrations of the pathway would be infinitely high. This is biologically unrealistic because there are factors that restrict intermediate concentrations, such as limited solvent capacity and osmotic constraints. Thus, it is often more meaningful to maximise the rate of a pathway under restrictions for enzyme and intermediate concentrations.¹²

6.3. Game-Theoretical Approaches to Studying Optimal Pathway Design

The above optimisation approaches offer a deeper insight into the evolutionary origin and advantages of properties of metabolic pathways. Simple optimisation is, however, not always sufficient for understanding evolutionary phenomena.²³ This is because selective forces depend on the ecological properties of the environment and its interplay with the evolving population. Changes in the properties of the evolving population may cause changes in the properties of the environment, which in turn changes the selective forces. This is particularly the case if the environment contains coevolving competitors that optimise their own strategies. The optimal use of metabolic resources may, for example, depend on how other competitors use the metabolic resource present in the environment. Considering the mutual interactions between properties of the evolving population and properties of the environment is essential for understanding more complex phenomena in the evolution of metabolism, such as the evolution of crossfeeding²⁴ or the cooperative use of energy resources.²⁵

In a crossfeeding interaction, two or more strains (or species) stably coexist on a single limiting resource. One of the strains grows on the primary resource but degrades it only partially and excretes a metabolite that serves as the resource of the second strain. The emergence of crossfeeding interactions has been observed in long-term evolution experiments on *E. coli* in chemostats with glucose as the limiting resource.^{17–19} The evolution of stable polymorphisms on a single limiting resource is not expected based on the competitive exclusion principle.²⁶ Therefore, it raises the question of what advantage two crossfeeding strains have over a single competitor that completely degrades the primary resource. Using game-theoretical simulations, we can show that crossfeeding may emerge as a consequence of the optimisation of three properties of ATP-producing pathways, namely maximisation of the rate of ATP production, minimisation of the enzyme concentrations and minimisation

of the intermediate concentrations. This stable co-existence of populations with different properties in their metabolism cannot be derived on the basis of simple optimisation approaches alone.

A further application of evolutionary game on the evolution of metabolism is the analysis of the consequences of trade-offs between rate and yield of ATP-producing pathways. As discussed above, these trade-offs arise from thermodynamic principles and from the presence of alternative pathways of ATP production with opposing properties in yield and rate such as fermentation and respiration. The existence of trade-offs between rate and yield raises the question of whether it is favourable to produce ATP at a high rate but low yield or at low rate but high yield. Using game-theoretical approaches we can show that fast ATP production with low yield can be seen as selfish resource use, while ATP production with high yield but at a low rate can be seen as cooperative resource use.²⁵ Furthermore, it can be shown that similar to other forms of cooperation, cooperative resource use is expected to evolve in spatially structured environments, while selfish resource use is expected to evolve in spatially homogeneous populations.

6.4. Genetic Robustness and Epistasis in Metabolic Pathways

In addition to offering explanations for the evolutionary origin of patterns of metabolism as the ones discussed above, an analysis of simple metabolic pathway models can help to derive predictions for phenomena related to pathway evolution such as genetic robustness and epistasis. Genetic robustness can be defined as robustness of fitness-relevant properties such as fluxes or steady-state metabolite concentrations against deleterious mutations of the enzymes. Genetic robustness can be quantified by a control coefficient C given by the ratio of the relative change of fitness and the relative change of a parameter,

$$C = \log(w/w^*) / \log(p/p^*),$$

where w/w^* is the ratio between the fitness of the perturbed and unperturbed system, and p/p^* is the ratio between the perturbed and unperturbed parameter. If, for example, a change in a parameter of 5% causes a 5% change in fitness, the control coefficient is one. Less robust systems – in which parameter changes result in larger fitness effects – are characterised by larger control coefficients; more robust systems are characterised by smaller control coefficients.

For small perturbations of a single reaction and if fitness is determined by a steady-state flux of a metabolic pathway, the above definition of robustness is equivalent to flux control coefficients in the framework of MCA.¹² Using MCA it can be shown that the flux control coefficients of all reactions over the flux of a pathway add up to one. In optimised pathways, the control over the flux is distributed over all enzymes of a pathway. This implies that the control coefficients are smaller than one, i.e., the changes in the flux of a pathway are smaller than the change in a parameter of a single enzyme. A similar line of reasoning applies to the evolution of

dominance.^{27,28} In these studies it is assumed that dominance corresponds to the loss of one functional allele and hence a reduction of gene expression by 50%. Such a reduction has a small effect when control coefficients are small. It has therefore been argued that dominance results as an intrinsic property of metabolic pathways.^{27,28}

In contrast to small deleterious mutations, the effects of complete knockouts of enzymes has not been studied in detail. This is because in simple models of metabolic pathways all enzymes are typically essential, i.e., a knockout of an enzyme leads to a steady-state flux of zero. However, in more complex networks, complete knockouts are not always lethal.^{29,30} Experimental findings and further theoretical details on robustness in large networks are discussed further below.

In addition to deriving predictions on the mutational robustness of metabolic pathways, MCA can also be used to derive predictions for the interactions between mutation. Interactions between mutations are described by epistasis. If the effect of two combined deleterious mutations is less severe than would be expected from the effect of each individual mutation, epistasis is positive; if it is more severe than expected, epistasis is negative. A common definition for epistasis is

$$e = w_{AB} - w_A w_B,$$

where w_{AB} , w_A and w_B are the relative fitness of the double mutant and the corresponding single mutants, respectively. Specific cases of epistatic interactions are compensatory mutations (the second mutation buffers the negative effects of the first mutation) and synthetic lethals (the double mutant is lethal although the two corresponding single mutants are viable).

Studies on interactions between mutations have recently received increasing interest. This is because interactions of mutations offer insights into the mechanistic interactions of the mutated compounds.³¹ Furthermore, epistasis is of fundamental importance for theories on the evolution of recombination and sexual reproduction.³²

On the basis of MCA, the following predictions for epistatic interactions in metabolic pathways can be derived. If an enzyme of an optimised pathway is affected by a deleterious mutation, it will typically get a higher control, i.e., it will become a stronger bottleneck for the flux compared to the unperturbed pathway. Since the control coefficients of all enzymes of a pathway add up to one, the control of the unaffected enzymes decreases. Therefore, a second mutation in the same enzyme will have a stronger effect than expected, i.e., epistasis is negative. A second mutation in a different enzyme typically has a smaller effect than expected, i.e., epistasis is positive. For small mutations, it can be shown that the mean of epistasis is zero.¹² The above line of reasoning is based on the assumption that the flux of a pathway is the only fitness-relevant property. Situations where other properties such as metabolite concentrations are relevant for the fitness of an organism have been described by Szathmáry.³³

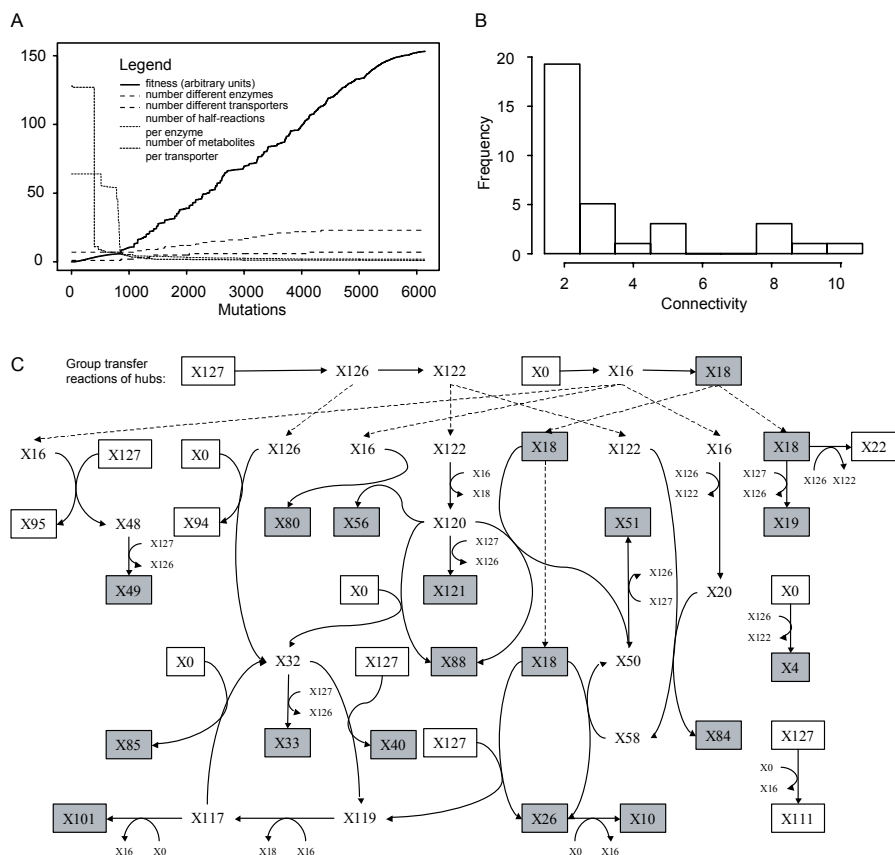


Fig. 6.1. Example simulation of the evolution of metabolic networks (reproduced from Ref. 43). (A) The initial network consists of 128 metabolites, seven unspecific enzymes (each of which transfers one of the seven biochemical groups that metabolites carry) and a single unspecific transporter. Within the course of evolution, the enzymes and transporter duplicate and increase in specificity (i.e., the number of half-reactions per enzyme and of metabolites per transporter decreases). The emerging network consists of 23 enzymatic reactions and seven transport processes. In the sample simulation, all enzymes and transporters in the emerging network are highly specific, i.e., the enzymes catalyse only two half-reactions and the transporters transport single metabolites. The emerging network contains only 33 metabolites. The remaining metabolites are not involved in the emerging network. (B) Connectivity distribution of the emerging group transfer network. Most metabolites are involved in only two reactions. However, a few metabolites are highly connected. (C) Pathway scheme of the emerging group transfer network. The metabolites X0 and X127 are taken up from the environment, whereas metabolites X4, X22, X94, X95 and X111 are excreted into the environment (white boxes). The network eventually transforms metabolites X0 and X127 into those metabolites that are involved in biomass formation (grey boxes). Interestingly, metabolite X4 is excreted although it is involved in biomass formation. Note that some half-reactions evolve, such as the one from X127 to X126, and monopolise the transfer of a specific group (in this case the first group in the binary string). These metabolites are involved in many reactions and therefore have high connectivity. The group transfer reactions of these hubs are summarised in the first line of the pathway scheme. The emerging group transfer network is much more complex than the corresponding monomolecular reaction network and even includes a cycle ($X32 \rightarrow X119 \rightarrow X117 \rightarrow X32$), with the net reaction of $X0 + X16 + X127 \rightarrow X18 + X40 + X85$). Further details of the simulation are given in the corresponding publication.⁴³

6.5. Large-Scale Properties of Metabolic Networks and Their Evolution

6.5.1. *Hubs and robustness in metabolic networks*

The theoretical studies presented above focus on the analysis of simplified models of metabolic pathways with comparably low complexity. The rapid increase in data on large metabolic networks in recent years allows the analysis of large-scale properties of metabolism from a network perspective. One such network property is the connectivity distribution. In metabolic networks, the connectivity refers to the number of reactions in which a given metabolite is involved. It has been reported that the connectivity distribution in metabolic networks follows approximately a power law.^{34,35} A power-law connectivity distribution implies that there are *hub* metabolites involved in a high number of reactions. Typical hub metabolites are ATP, NADH, glutamate, coenzyme A and their derivatives. Interestingly, these metabolites often play a key role in the transfer of biochemical groups.

One possible mechanism by which power-law connectivity distributions may emerge in growing networks is the preferential attachment of new nodes to existing ones with high connectivity.³⁶ Mechanisms such as preferential attachment are typically based on the assumption that selection acts on individual nodes or edges. These mechanisms, however, do not consider that in biochemical reaction networks fitness is determined by the properties of the entire network rather than its components. Therefore it is questionable whether preferential attachment is applicable to the evolution of metabolic networks.

Some authors have suggested that the benefits of power-law connectivity distributions may arise from network robustness.^{34,37} However, whether robustness is a strong selective force in the evolution of metabolic networks is questionable. First, theoretical considerations suggest that the evolution of genetic redundancy (a form of robustness against knockouts) only works under very specific conditions in terms of mutation rates, gene functions and interactions.^{38,39} Second, a recent study on robustness and enzyme indispensability in yeast metabolism indicates that the apparent dispensability of many enzymes is not due to network robustness but the fact that many enzymes are only required under specific environmental conditions.³⁰ Third, robustness against environmental changes is also unlikely to explain the connectivity distributions observed in natural networks. This is because power-law connectivity distributions have been observed in a wide range of organisms living in very different environments, including, for example, intercellular parasites that may live in very stable environments.⁴⁰ Finally, no evolutionary scenarios have been presented to demonstrate that selection for increased robustness leads to the emergence of metabolic networks with power-law connectivity.

A number of alternative scenarios for the evolution of genetic robustness that do not rely on direct selection have been proposed.³⁹ Specifically it has been ar-

gued that genetic robustness may be an intrinsic property of specific systems. As described above, this scenario has been supported on the basis of MCA at least for small deleterious mutations and for dominance. An alternative explanation is that robustness against deleterious mutations may emerge as a side product of selection for robustness against environmental perturbations. This view is supported by observations that many knockouts are viable because the corresponding enzyme is not required in the given experimental conditions.³⁰

6.5.2. *Computer simulations of scenarios for the evolution of metabolism*

To study the evolution of robustness and the emergence of hubs in metabolic networks we implemented computer simulations of a widely accepted evolutionary scenario originally proposed by Kacser and Beeby.⁴¹ According to this scenario complex metabolic networks characterised by large numbers of enzymes with high specificity evolved from ancestral networks consisting of few enzymes with broad specificity. The broad specificity allowed all essential metabolic functions to be maintained at the cost of low rate constants for any single biochemical reactions. Networks were selected for growth rate and evolved by mutations affecting the kinetic properties of the enzymes and occasional gene duplications. Although a number of alternative scenarios for the evolution of novel enzymes and metabolic pathways have been proposed,⁴² this scenario is a plausible mechanism for the early evolution of metabolic networks. An example simulation is shown in Fig. 6.1. Based on our simulations we can confirm that this scenario indeed leads to the emergence of metabolic networks with connectivity distributions similar to those observed in nature if important biochemical constraints are incorporated.⁴³ In particular, we can show that hubs emerge only in group transfer networks. Hubs emerge because some metabolites *monopolise* the transfer of specific groups. This is in line with the observation that most hubs in natural networks such as ATP or NADH are key players in the transfer of biochemical groups. Our scenario indicates that hubs emerge in the network as a consequence of selection for growth rate. Therefore, direct selection for robustness is not required to explain the emergence of hubs in metabolic networks.

6.5.3. *Robustness and epistasis in the emerging networks*

Figure 6.2 shows the effect of mutations on the networks emerging in the simulation. The effects of small deleterious mutations of the enzymes on the flux of the emerging networks are comparably small, i.e. all control coefficients are close to zero, see Fig. 6.2A. Thus the emerging networks are robust against slightly deleterious mutations that affect the enzymes. In contrast, a large fraction of complete knockouts of enzymes is lethal, see Fig. 6.2B. Thus, the emerging networks are not robust against complete knockouts of enzymes. However, the emerging networks

contain a few enzymes that are beneficial but non-essential to the functioning of the network. The relative fitness of knockouts of these non-essential enzymes is distributed approximately uniformly between 0 and 1.

Figure 6.2C and Fig. 6.2D show the distribution of epistasis for small mutations and complete knockouts of enzymes, respectively. Epistasis of small deleterious mutations follows an asymmetric distribution with a mean close to zero and a positive median. Most interactions between mutations are characterised by small positive epistasis. On the other hand, there are mutations characterised by comparably large negative epistasis. As described above, this is because the first mutation results in an increased control of the affected enzyme, and in a decreased control of all other enzyme. Epistasis between complete knockouts of enzymes follows a different pattern. Because epistasis is zero if the double mutant and at least one single mutant are lethal, we include only those interactions where either both single mutants, or the double mutant is viable.

The distribution of epistasis is characterised by a positive mean and a positive median. Two mutations that knock out the function of the same enzyme always have positive epistasis (if the knockout is viable). This is because the double mutant has the same fitness as the single mutants. A second mutation that knocks out an enzyme that is already non-functional because of the first mutation has no further effect on fitness. This is in contrast to small deleterious mutations where two mutations that affect the same enzyme always have negative epistasis.

6.6. Conclusion

Metabolic networks are ideally suited for theoretical analyses because they are perhaps the best studied network type in biology. In contrast to signal transduction or gene regulation networks, typically all participating components are known. Although there is only limited data, the kinetics of metabolic networks is still better characterised than other types of networks. Moreover, the mathematical theory of metabolism is very well developed. Combining this theory with approaches from evolutionary biology helps the understanding of a wide range of patterns observed in cellular metabolism.

Many properties of large metabolic networks can be derived from theory and from approaches to simplified systems with comparably low complexity. The high robustness of metabolism towards small deleterious mutations of the enzymes as well as the distribution of epistatic effects between these mutations result from intrinsic properties of metabolism. This is supported by our studies on the evolution of large metabolic networks, which result in conclusions in line with findings derived from relatively simple metabolic pathway models.

However, some properties of metabolic networks such as their connectivity distribution or their robustness towards complete knockouts of enzymes require theoretical approaches using complex network models. Using computer simulations

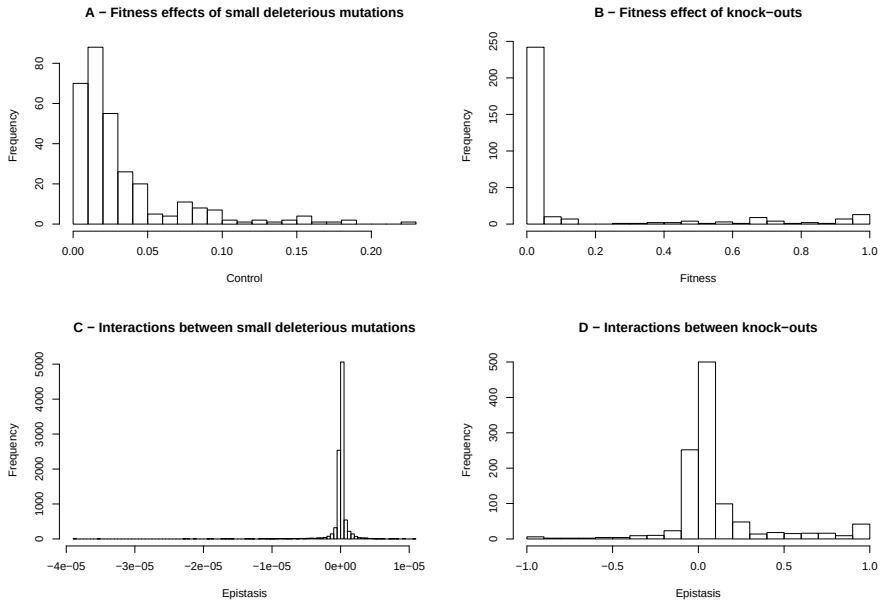


Fig. 6.2. Robustness and epistasis in the emerging metabolic networks. The histograms show the effect of mutations in 10 networks emerging in the simulations presented in Ref. 43. (A) The robustness of the biomass formation of the networks towards small deleterious mutations in the enzymes or transporters is quantified using control coefficients. The control coefficients quantify the relative response of the rate of biomass formation (which is proportional to fitness in the simulations) towards the small change in the activity of an enzyme or transporter. The figure shows that the control coefficients are close to zero. This implies that the networks are robust towards small changes in the activity of the enzymes, i.e. the network is robust against small deleterious mutations. (B) Robustness towards complete knockout of enzymes or transporters. The histogram shows the distribution of the relative fitness values after complete knockout of an enzymatic reaction or transport process. Most knockouts have a fitness of zero, i.e., are lethal. However, the networks contain a few non-essential biochemical reactions. (C) Epistasis between small deleterious mutations. The distribution of epistatic interactions is asymmetric. It has an average close to zero and a positive median. This is because mutations that affect the same enzyme have comparably strong negative epistasis, while mutations that affect different enzymes tend to have small positive epistasis. (D) Epistasis between viable knockouts. The distribution shows only those interactions where either both single mutants or the double mutant are viable. In the other cases, epistasis is zero. The distribution between has a positive average and a positive median. In contrast to small deleterious mutations, viable knockouts that affect the same enzyme always have positive epistasis. This is because the single mutant has the same fitness as the double mutant. A second mutation that knocks out a function that has already been disrupted by the first mutation has no fitness effect.

to study scenarios of the evolution of comparably large metabolic networks allows insights to be gained into the emergence of hub metabolites. These simulations indicate that hubs may emerge as a consequence of selection for growth rate. Direct selection for robustness is not required to explain the emergence of hubs in

metabolic networks.

Although the emerging networks have high robustness towards small deleterious mutations, they have low robustness against complete knockouts of enzymes. This is in contrast to the observation that many enzymes are dispensable.³⁰ However, this high robustness arises mainly because most enzymes are only required under specific environmental conditions. To study the relation between environmental robustness and genetic robustness, the approaches presented above can be extended to account for selection in variable environments.

The examples discussed here demonstrate that mathematical approaches combined with evolutionary theory have considerable potential to develop a better understanding of generic properties of metabolic networks. In future these approaches may usefully be extended to study the design of other biochemical reaction networks such as signal transduction or gene regulation.

References

1. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, The KEGG resource for deciphering the genome, *Nucleic Acids Research*. **32**, D277–280, (2004).
2. J. Papin, J. Stelling, N. Price, S. Klamt, S. Schuster, and B. Palsson, Comparison of network-based pathway analysis methods, *Trends in Biotechnology*. **22**, 400–405, (2004).
3. J. S. Edwards and B. O. Palsson, The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities, *Proceedings of the National Academy of Science USA*. **97**, 5528–5533, (2000).
4. J. Forster, I. Famili, P. Fu, B. Palsson, and J. Nielsen, Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network, *Genome Research*. **13**, 244–253, (2003).
5. S. Becker and B. O. Palsson, Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation, *BMC Microbiology*. **5**, 8, (2005).
6. J. L. DeRisi, V. R. Iyer, and P. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*. **278**, 680–686, (1997).
7. B. H. ter Kuile and H. V. Westerhoff, Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway, *FEBS Letters*. **500**, 169–171, (2001).
8. M. K. Oh, L. Rohlin, K. C. Kao, and J. C. Liao, Global expression profiling of acetate-grown *Escherichia coli*, *Journal of Biological Chemistry*. **277**, 13175–13183, (2002).
9. O. Fiehn, Metabolomics and the link between genotypes and phenotypes, *Plant Molecular Biology*. **48**, 155–171, (2002).
10. U. Sauer, High-throughput phenomics: experimental methods for mapping fluxomes, *Current Opinion in Biotechnology*. **15**, 58–63, (2004).
11. I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg, BRENDA, the enzyme database: updates and major new developments, *Nucleic Acids Research*. **32**, D431–433, (2004).
12. R. Heinrich and S. Schuster, *The regulation of cellular systems*. (Chapman & Hall, New York, NY, 1996).
13. H. Bialy, Living on the edges, *Nature Biotechnology*. **19**, 111–112, (2001).

14. D. A. Fell, Metabolic control analysis: a survey of its theoretical and experimental development, *Biochemical Journal*. **286**, 313–330, (1992).
15. D. E. Dykhuizen and A. M. Dean, Enzyme activity and fitness: Evolution in solution, *Trends in Ecology and Evolution*. **5**, 257–262, (1990).
16. R. E. Lenski and M. Travisano, Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations, *Proceedings of the National Academy of Science USA*. **91**, 6808–6814, (1994).
17. R. B. Helling, Speed versus efficiency in microbial growth and the role of parallel pathways, *Journal of Bacteriology*. **184**, 1041–1045, (2002).
18. R. F. Rosenzweig, R. R. Sharp, D. S. Treves, and J. Adams, Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*, *Genetics*. **137**, 903–917, (1994).
19. S. Treves, D. S. Manning and J. Adams, Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*, *Molecular Biology Evolution*. **15**, 789–797, (1998).
20. S. S. Fong and B. O. Palsson, Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes, *Nature Genetics*. **36**, 1056–1058, (2004).
21. T. G. Waddell, P. Repovic, E. Meléndez-Hevia, R. Heinrich, and F. Montero, Optimization of glycolysis: a new look at the efficiency of energy coupling, *Biochemical Education*. **25**, 204–205, (1997).
22. A. Stephani, J. C. Nuno, and R. Heinrich, Optimal stoichiometric designs of ATP-producing systems as determined by an evolutionary algorithm, *Journal of Theoretical Biology*. **199**, 45–61, (1999).
23. T. Pfeiffer and S. Schuster, Game-theoretical approaches to studying the evolution of biochemical systems, *Trends in Biochemical Sciences*. **30**, 20–25, (2005).
24. T. Pfeiffer and S. Bonhoeffer, Evolution of crossfeeding in microbial populations, *American Naturalist*. **163**, E126–135, (2004).
25. T. Pfeiffer, S. Schuster, and S. Bonhoeffer, Competition and cooperation in the evolution of ATP-producing pathways, *Science*. **292**, 504–507, (2001).
26. G. Hardin, The competitive exclusion principle, *Science*. **131**, 1292–1297, (1960).
27. H. Kacser and J. E. Burns, The molecular basis of dominance, *Genetics*. **97**, 639–666, (1981).
28. L. D. Hurst and J. P. Randerson, Dosage, deletions and dominance: Simple models of the evolution of gene expression, *Journal of Theoretical Biology*. **205**, 641–647, (2000).
29. J. Stelling, S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles, Metabolic network structure determines key aspects of functionality and regulation, *Nature*. **420**, 190–193, (2002).
30. B. Papp, C. Pal, and L. D. Hurst, Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast, *Nature*. **429**, 661–664, (2004).
31. A. H. Tong and et al., Global mapping of the yeast genetic interaction network, *Science*. **303**, 808–813, (2004).
32. N. H. Barton and B. Charlesworth, Why sex and recombination?, *Science*. **281**, 1986–1990, (1998).
33. E. Szathmáry, Do deleterious mutations act synergistically? Metabolic control theory provides a partial answer, *Genetics*. **133**, 127–132, (1993).
34. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi, The large-scale organization of metabolic networks, *Nature*. **407**, 651–654, (2000).
35. A. Wagner and D. A. Fell, The small world inside large metabolic networks, *Proceedings of the Royal Society London, Series B Biological Sciences*. **268**, 1803–1810,

- (2001).
36. A. L. Barabasi and R. Albert, Emergence of scaling in random networks, *Science*. **286**, 509–512, (1999).
 37. R. Albert, H. Jeong, and A. L. Barabasi, Error and attack tolerance of complex networks, *Nature*. **406**, 378–382, (2000).
 38. M. A. Nowak, M. Boerlijst, J. Cooke, and J. Smith, Evolution of genetic redundancy, *Nature*. **388**, 167–171, (1997).
 39. J. de Visser, J. Hermisson, G. Wagner, L. Meyers, H. Bagheri-Chaichian, J. Blanchard, L. Chao, J. Cheverud, S. Elena, W. Fontana, G. Gibson, T. Hansen, D. Krakauer, R. Lewontin, C. Ofria, S. Rice, G. von Dassow, A. Wagner, and M. Whitlock, Evolution and detection of genetic robustness, *Evolution*. **57**, 1959–1972, (2003).
 40. H. Ma and A. P. Zeng, Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms, *Bioinformatics*. **19**, 270–277, (2003).
 41. H. Kacser and R. Beeby, Evolution of catalytic proteins or on the origin of enzyme species by means of natural selection, *Journal of Molecular Evolution*. **20**, 38–51, (1984).
 42. S. Schmidt, S. Sunyaev, P. Bork, and D. T., Metabolites: A helping hand for pathway evolution, *Trends in Biochemical Sciences*. **28**, 336–341, (2003).
 43. T. Pfeiffer, O. Soyer, and S. Bonhoeffer, The evolution of connectivity in metabolic networks, *PLoS Biology*. **3**, e228, (2005).

Chapter 7

Protein Interactions from an Evolutionary Perspective

FLORENCIO PAZOS¹ AND ALFOSO VALENCIA²

¹ *Computational Systems Biology Group, National Centre for Biotechnology (CNB-CSIC), Spain*

² *Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Spain*
pazos@cnb.csic.es, valencia@cnb.uam.es

Interpreting the massive amounts of available genomic information in functional terms requires, among other things, discernment of the interactome determined by a given proteome. To accomplish this task, experimental techniques for the high-throughput determination of sets of interacting proteins can be assisted by computational approaches. These approaches, in spite of having their own limitations and problems, can overcome some of the intrinsic drawbacks associated with the experimental techniques including the error associated with the high-throughput determination of protein interactions. Moreover, the computational approaches are comparable to their experimental counterparts in terms of accuracy. Because of the complexity in detecting interaction partners based on basic principles (using solely the physico-chemical features of the proteins), current computational methods look for interaction partners by searching for the trail that the process of adaptation to specific interactors leaves in the sequences and genomic features during the evolutionary process.

7.1. Introduction

Paradoxically, one of the main realizations of the so called post-genomic era is that the genetic repertoires of the organisms can not account for many of their complex characteristics or for the differences between the organisms themselves (neither the number of genes nor their characteristics). Consider, for example, the similar number of genes between the plant *Arabidopsis thaliana* and human, or the almost identical genes of mouse and human. Since the protein repertoires of very different organisms are unexpectedly similar, the differences should arise from higher levels of complexity. Biological systems are the prototype of complex systems, where the whole is more than the sum of its parts.^{1–3} By only considering the complex network of relationships between cellular components we can go one step further to understand many of the features characterizing living systems. In the case of proteins, the basic functional and structural units of cellular systems, it is becom-

ing clear that their individual functions cannot account for many properties of the system at higher levels, and only in the context of their interactions and complex relationships with others are their functions realized in biological terms. This is why it is very important to decipher the interactome for a given proteome. This interactome, the network of protein-protein interactions of a given organism, contains essential information about its biology because protein interactions are involved in most cellular processes: macromolecular complexes, signalling cascades, metabolism (interaction between consecutive enzymes in metabolic pathways), transcriptional control, etc. This importance of deciphering interactomes has led to the development of techniques for the massive determination of protein interactions (Uetz and Finley, 2005), such as the yeast two-hybrid system⁴ or affinity purification of complexes followed by mass spectrometry analysis.^{5,6} These techniques were applied in a high-throughput way aiming to determine as much as possible of a given interactome. They were used to determine large proportions of the interactomes of a number of model organisms, ranging from bacteria such as *H. pylori*⁷ or *E. coli*⁸ to human,⁹ covering unicellular eukaryotes like yeast^{5,6,10,11} or multicellular organisms like *C. elegans*¹² or *D. melanogaster*.¹³ These first high-throughput experimentally determined proteomes still contain a considerable degree of error^{14–16} when assessed in terms of individual pairs of interacting proteins. It can be said that they provide an overall view of the complete interactome and its properties (see below) at the expenses of losing accuracy in terms of individual interactions. This is a feature common to other high-throughput techniques such as DNA arrays, where overall pictures of the expression of genomes are obtained at the cost of dealing with errors in the expression levels of individual genes.^{17,18}

Knowledge of these first (still incomplete) interactomes allowed for some of the first studies of biological networks from a systems biology point of view, extracting important data on the topology, connectivity, evolution and functionality of global protein interaction networks.^{19–25}

Computational approaches can complement these experimental methods on many different levels. Computational techniques are behind most of the global studies of the interactome discussed in the previous paragraph since they involve handling huge amounts of data. They are also implicated in the efficient representation and storage of the evolving datasets related with protein interactions.²⁶ But more importantly, they are at the base of the determination of protein interactions itself. Computational approaches can be used to guide experiments by restricting the number of pairs to test experimentally instead of blindly trying all against all,²⁷ to filter the intrinsically noisy experimental interactions and to combine them with other information in order to increase the accuracy,²⁸ or to predict interactions purely *in silico*.

Most of the methods for the *in silico* prediction of interacting proteins are directly or indirectly based on evolutionary features. The tremendous complexity of the protein-protein interaction phenomena, including the existence of different

types of complexes (transient, permanent), the low interaction energy of the complex, the uncertain dependence on a small number of key residues (*hot spots*), etc., makes almost intractable the *ab initio* prediction of interaction partners (based solely on their sequences and/or structures).^{29–32} On the other hand, we can obtain information on interacting pairs of proteins by comparative genomics, looking for their evolutionary landmarks, since interacting proteins are expected to present particular evolutionary features (mainly coevolution).

This review tries to give an overview of the current landscape of computational techniques for predicting pairs of interacting proteins from sequence and/or genome information, focusing on the ones based on evolutionary information. Methods for predicting protein regions involved in interaction, docking methods and others are not included in this article and they are covered in other excellent reviews.^{30,32–34}

7.2. Computational Prediction of Protein Interactions

7.2.1. *Experimental vs. computational methods*

As discussed in the introduction, experimental methods for the high-throughput determination of protein interactions have a high degree of error when evaluated in terms of individual pairs.^{14–16} For example, the intersection between the three sets of interacting pairs detected in three independent experiments, in which yeast two-hybrid was used to massively determine interaction partners in yeast was only of 6 pairs³⁵ and the accuracy of these approaches was estimated to be as low as 10%.¹⁶ In spite of this low accuracy and the amazingly lack of agreement between experiments when assessed in terms of pairs, the global characteristics of the interaction networks are quite similar (scale-free topology, hubs, etc.) which justifies the utility of these networks for global studies.³⁶ Another drawback of these high-throughput experimental techniques is the low coverage. These approaches are still far from being truly high-throughput, in the sense that the intrinsic drawbacks of the methodology allow only a fraction of all possible pairs of proteins to be tested.³⁵ Other limitations of these techniques, consequences of their experimental nature, include the tendency to preferentially detect interactions between highly expressed proteins or between proteins belonging to some cellular compartments to the detriment of others.¹⁶

These drawbacks of the high-throughput experimental techniques for the determination of sets of interacting proteins further justified the development of computational methods to complement them. Computational methods for the prediction of protein interactions have been shown to have similar (or even higher) level of accuracy than experimental ones when combined under certain circumstances.¹⁶ Moreover they are cheaper and faster than their experimental counterparts and do not share the same limitations, like being influenced by the abundance of proteins or their cellular compartment (see above). These methods are based on simple genomic or sequence features intuitively related to interaction (Fig. 7.1), such as

conservation of gene neighbouring across genomes, domain fusion events, comparison of phylogenetic distributions (patterns of presence/absence of genes in a set of genomes), correlated mutations and similarity of phylogenetic trees, among others.

7.2.2. *Conservation of gene neighbouring*

One of the simplest evolutionary features related to interaction one can look for is the closeness of interacting partners in the genome, and the conservation of this closeness across distant organisms. The idea behind it is that interacting or, in general, functionally related proteins are close in a genome in order to allow joint transcriptional control. This is especially clear in prokaryotic organisms, where operons (sets of contiguous genes sharing a promoter and hence under the same transcriptional control) are widespread. In eukaryotic organisms this way of controlling transcription using operons is not common and consequently the tendency of functionally related genes to be close in the genome is not so evident. This neighbourhood relationship is more meaningful when it is conserved in distant species,³⁷ since in close species the genomic context of a gene may be conserved just because of the short divergence time. So although at first sight it seems trivial to detect these conserved pairs of close genes, the actual methods involve a number of parameters to tune, like the chromosomal distance between the two genes and the phylogenetic distance between the species.^{38,39} The basic gene neighbourhood methodology to predict if two proteins A1 and B1 in organism 1 are functionally related consist of: (i) Evaluating whether A1 and B1 are close in genome 1 according with some genomic distance cutoff, (ii) looking for their corresponding orthologues in another organism (A2, B2), using for example the BLAST best bi-directional hit method, (iii) applying to A2-B2 the same distance cutoff, (iv) eventually, repeating steps (ii) and (iii) with other distant organisms in order to assess whether this neighborhood relationship is conserved in more organisms (A3-B3, A4-B4, etc.) (Fig. 7.1B).

These methods have been used to locate a number of pairs of physically or functionally related proteins the prototypical case being the Tryptophan operon, whose members are close in a number of phylogenetically distant bacteria.^{38,39} The obvious drawback of this technique is its limitation of using bacterial genomes as a source of information, where there is a clear tendency to put together functionally related genes in operons. This makes it impossible to apply the technique to proteins typical of eukaryotic organisms (without homologues in prokaryotes).

7.2.3. *Gene fusion*

A gene fusion event is detected when two independent proteins in a given organism(s) are fused as two domains of the same polypeptide (and hence coded by the same gene) in another organism(s) (Fig. 7.1C). Since in the second case it is clear that the two domains are interacting and involved in the same function, it is reasonable to conclude that the homologues of these domains, which are in separate

polypeptides in the first case, are going to be involved in the same function too. Enright *et al.*⁴⁰ and Marcotte *et al.*⁴¹ developed algorithms to detect such fusion events in genomic sequences. The basic algorithm is simply based on detecting pairs of proteins in a given organism which share sequence similarity (BLAST) with the same protein in another organism, which would indicate a possible fusion event. An obvious problem of the described approach is that modular domains present in a high number of proteins would produce false positives. For example, all proteins with SH3 domains would be predicted to interact with each other. One way of overcoming this is to exclude similarities due to these domains, or (*a posteriori*) to exclude from the list of predicted interactions the ones involving promiscuous proteins (proteins predicted to interact with too many others).

Marcotte *et al.*⁴¹ proposed an evolutionary hypothesis for explaining such fusion events: if two proteins A and B have to interact in order to perform a given function, the concentration of the active complex would be much higher if the two proteins are fused together than if the two proteins are separated and hence rely on Brownian motion to find each other and form the active complex.

Examples of domain fusions include the *E. coli* histidine biosynthesis proteins HIS2 and HIS10, which are fused in yeast in one single polypeptide (HIS2) with two domains clearly homologous to the two *E. coli* proteins.⁴¹ It has indeed been shown that metabolic proteins are frequently involved in domain fusion events.⁴² One advantage of this approach for detecting protein associations is its reliability, since the fact that two proteins are fused is a clear indication of their functional relationship (except for promiscuous domains, see above). Hence, this approach produces almost no false positives. Its disadvantage is its range of applicability because these fusion events, while very informative, are not very frequent, especially in prokaryotes. For example, Enright *et al.*⁴⁰ detected only 64 unique fusion events in 3 bacterial complete genomes.

7.2.4. Similarity of phylogenetic profiles

A phylogenetic profile is a pattern of presence/absence of a given protein in a set of organisms. It represents the species distribution of that protein (Fig. 7.1D). Their utility in predicting protein interactions and functional relationships comes from the fact that pairs of interdependent proteins tend to have similar phylogenetic profiles. That is, the two proteins tend to be present in the same subset of organisms and absent together in the complementary set.^{41,43,44} The idea behind this approach is that proteins which need each other to perform a given function will be either both present or both absent. In the second case this is due to reductive evolution because the organism (especially bacteria) would get rid of one of the genes if the other required partner is not present.

In the first versions of the phylogenetic profile methodology for predicting interactions, the species distribution of a protein was represented qualitatively, as a binary vector where 1 coded for the presence of that protein in an organism and 0

for its absence (Fig. 7.1C). In that case, the similarity of phylogenetic distributions was evaluated as the distance between these binary vectors (e.g. Hamming distance or mutual information). If PA and PB are the binary phylogenetic profiles of two proteins A and B, where PA_i codes for the presence of protein A in the genome i th of a set of n genomes (1 if it is present and 0 otherwise, according to a given criteria of orthology), the Hamming distance is defined as

$$d_{AB} = \sum_{i=1}^n |PA_i - PB_i|.$$

This distance represents the number of different bits between the two profiles or, in other words, the number of organisms where one protein is present and the other absent or *vice versa*. It was shown that similar vectors (low distance) were related with real interaction partners.⁴⁴ Later, quantitative information was incorporated by encoding in the positions of the vector the BLAST⁴⁵ E-value of a protein in a given organism with respect to an organism of reference.⁴⁶ In this case, mutual information⁴⁷ is used to calculate the distance between two vectors after discretizing their values. In this way, not only the presence/absence of the protein is taken into account but their phylogenetic distances, to some extent, as well. In this case, the i th position of the phylogenetic profile for protein A, instead of being just 1 or 0, is calculated as

$$PA_i = -1/\log(E_{Ai})$$

where E_{Ai} is the E-value of protein A in organism i with respect to an organism of reference. Values of $PA_i > 1$ are truncated to 1. From these vectors, the mutual information between the phylogenetic profiles of proteins A and B is calculated as

$$MI(A, B) = -\sum p(a) \ln(a) - \sum p(b) \ln(b) + \sum \sum p(a, b) \ln(p(a, b))$$

where $p(a)$ and $p(b)$ are the binned distribution of PA_i and PB_i values respectively (for example, in 0.1 intervals) and $p(a, b)$ the corresponding joint probability distribution. The sums run for all the bins in the distributions. The relationship between the power of this methodology for detecting interacting pairs of proteins and its parameters (E-value cutoff, number and phylogeny of the set of organisms for constructing the profiles, etc.) has been studied.^{48,49}

Not only similar profiles are informative but also anti-correlated ones (one protein is present when the other is absent and *vice versa*). These anti-correlated profiles have been related with enzyme displacement in metabolic pathways.⁵⁰ Furthermore, this versatile technique has recently been extended to triplets of proteins, allowing the search for more complicated patterns of presence/absence (e.g. protein C is present if A is absent and B is also absent). This allows the detection of interesting cases representing biological phenomena beyond binary functional interactions, like complementation.⁵¹

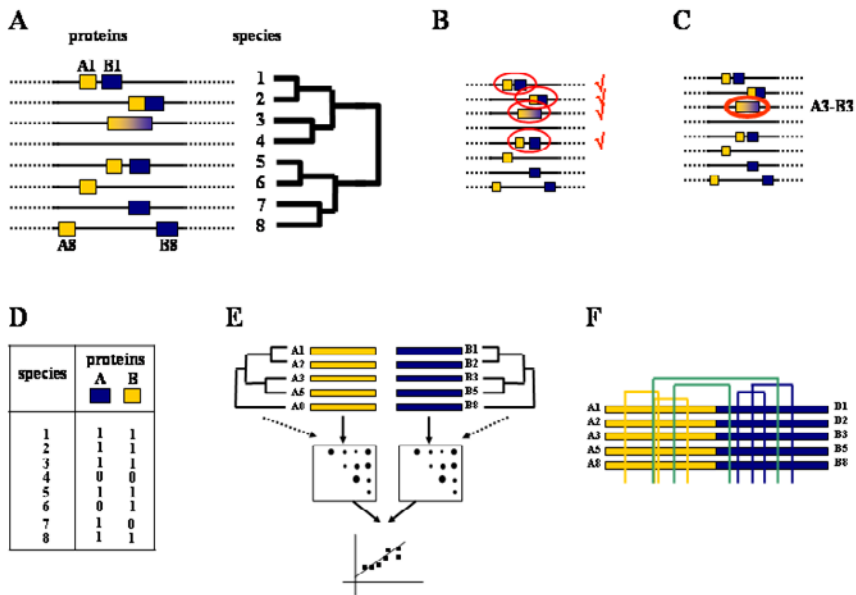


Fig. 7.1. Evolution-based methods for assessing the possible interaction between two proteins. (A) Sequence and genomic information about two proteins (A and B, yellow and blue) is used to assess their possible interaction. The sequences and genome positions of the orthologs of the two proteins (A1...A8, B1...B8) in a number of organisms related by a phylogeny (1...8) are used. (B) *Conservation of Gene Neighbouring*. The number of genomes where both proteins are close (genomes 1, 2, 3 and 5 in this example) and their phylogeny are used to assess whether the proteins are interacting or not. (C) *Gene Fusion*. Genomes are sought where both proteins appear as part of a single polypeptide (species 3 in this example). (D) *Similarity of Phylogenetic Profiles*. Phylogenetic profiles of both proteins are constructed by assessing the presence (1) or absence (0) of the two proteins in the set of species, and the similarity between these profiles is evaluated. (E) *Similarity of Phylogenetic Trees* (mirror-tree). Multiple sequence alignments for the two proteins are built. Only sequences coming from organisms where both proteins are present are used (genomes 1, 2, 3, 5 and 8 in this example). These multiple sequence alignments are used to generate distance matrices for both sets of orthologues. Alternatively, these multiple sequence alignments can be used to generate the actual phylogenetic trees and the distance matrices extracted from them. The similarity of these distance matrices is used as an indicator of interaction. Eventually, the phylogenetic distances between the species involved can be incorporated into the method for correcting the background similarity expected between the trees due to underlying speciation events and/or to detect non standard evolutionary events. (F) *Correlated Mutations*. The same multiple sequence alignments as in mirror-tree are used here to calculate intra- and inter-protein correlated mutations. The distributions of correlation values in these three sets are used to calculate an interaction index between the two proteins.

One disadvantage of this approach is that it can only be applied to complete genomes (as only then is it possible to be sure of the absence of a given gene). Similarly, it cannot be used with the essential proteins that are common to most organisms since these would be represented by profiles with 1 in all the positions and hence be without enough information.

7.2.5. *Similarity of phylogenetic trees*

Another coevolution-based method for detecting interaction partners is the one based on the detection of similar phylogenetic trees (Fig. 7.1E). It has been already qualitatively shown for some examples of interacting families of proteins, like insulin and its receptors⁵² or dockerins and cohexins,⁵³ that the phylogenetic trees of these interaction partners are more similar than expected. Possible explanations for explaining this similarity are that interacting proteins bear a similar evolutionary pressure (since they are involved in the same cellular process), and that they are forced to adapt to each other, both factors resulting in similar evolutionary histories. This coevolution between interacting proteins has been observed not only at the sequence level but also in other features like gene expression.⁵⁴

This similarity between phylogenetic trees of interacting proteins qualitatively observed was later quantified and tested in large datasets of proteins and protein domains^{55,56} statistically showing its capacity for detecting interacting pairs of proteins. This mirror-tree approach for predicting interactions is based on the comparison of protein distance matrices (using a linear correlation coefficient) instead of phylogenetic trees themselves (Fig. 7.1E). The exact comparison of phylogenetic trees is a complex and partially unsolved problem, and the direct comparison of distance matrices has been shown to be a convenient shortcut very useful in the special case of detecting protein interactions. So, for two proteins A and B with n species in common in their multiple sequence alignments, dA_{ij} being the distance between species i and j in the tree of protein A and dB_{ij} the corresponding distance in the tree of protein B, the similarity between their evolutionary histories (r_{AB}) is calculated as

$$r_{AB} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (dA_{ij} - \overline{dA}) (dB_{ij} - \overline{dB})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (dA_{ij} - \overline{dA})^2} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (dB_{ij} - \overline{dB})^2}},$$

where $\overline{dA}$ and $\overline{dB}$ are the average values of the corresponding distances. As a measure of distance between two proteins, the first versions of the method used the average sequence similarity extracted from the multiple sequence alignment.⁵⁶ Subsequent improvements of the method used distances directly extracted from the phylogenetic trees.⁵⁷

This simple and intuitive mirror-tree methodology has been applied to many proteins, and different implementations and variations of it have been developed.^{57–68} Ramani & Marcotte used this concept of similarity of trees to look for the correct mapping between two families of interacting proteins (e.g. to choose which ligand within a family interacts with which receptor within other families). The idea is that the correct mapping (set of relationships between the leaves of both trees) will be the one maximizing the similarity between both trees.⁶⁵

Another obvious extension of the method has been to incorporate information on the phylogeny of the species involved in the trees.^{57,67} The reason is that any pair of

trees is expected to have a background similarity due to the underlying speciation process, regardless the interaction of the corresponding proteins. It was shown that correcting by these background distances between species considerably increases the predictive power of the method.^{57,67} The correction is done either by using the phylogenetic distances between species taken from the standard tree-of-life based on an accepted molecular marker, the 16SrRNA,^{57,67} by averaging the values of the distance matrices, or by analyzing the principal components of these matrices.⁶⁷ The method by Pazos *et al.* allows also non-standard evolutionary events like horizontal gene transfers (HGT) to be detected, concomitantly with the prediction of interactions, since the 16SrRNA tree is used not only to correct the protein distances but also to assess whether they follow the standard phylogeny it symbolizes or not. Detecting those HGT cases is important in evolution-based interaction prediction methods because these proteins, due to their special evolutionary histories, do not fulfil some of the assumptions of many of these methods (like vertical inheritance). It has indeed been shown that excluding these automatically detected HGT cases from the predictions improves the performance.⁵⁷

The performance of this methodology has also been recently improved by using information on the coevolutionary context of a given pair of proteins.⁶² In this technique, the whole network of pairwise coevolutions within a genome is used to reassess the significance of a given coevolutionary signal. To conclude that two proteins A and B are coevolving, not only their isolated pairwise co-evolution r_{AB} is used (see above), but the similarity of their coevolutionary behaviours with the rest of the proteome, that is, the correlation between the vectors containing all the pairwise coevolutions for these two proteins (r_{Ai} and r_{Bi}) is also calculated.⁶²

The coevolution of interacting proteins is not only evident at the whole-sequence level but at sub-protein levels as well. It has been shown that this similarity of distance matrices between interacting proteins is more evident when its calculation is restricted to the residues forming the actual interaction surfaces, instead of using the full sequences of the proteins.⁶⁹ It looks like the co-evolutionary signal is also evident between protein domains, so that phylogenetic trees constructed for individual domains can be used to detect the domains actually involved in the interaction between two interacting multidomain proteins.⁷⁰

The obvious disadvantage of this method is the need for large numbers of homologous sequences to construct the trees. Moreover, the last versions of the method use the phylogenetic trees of a whole proteome, and hence require reliable protocols for the automatic and fast generation of these trees on a genomic scale.

7.2.6. *Correlated mutations*

When proteins belonging to the same family are aligned and equivalent residues are compared, some pairs of positions show a concerted mutational behavior, meaning that the amino acid changes in one position are related to the changes in the other. It has been shown that these pairs of positions are weakly related to spatial close-

ness between the corresponding residues in the three-dimensional structure of the protein.^{71,72} The underlying hypothesis for explaining such a relationship involves compensatory changes in one position to accommodate changes in the other. When this concept of correlated mutations was extended to inter-protein pairs of positions (one of the positions belonging to one protein/domain and the other to a different one) it was shown that these inter-protein correlated pairs tend to point to the interaction surface.⁷³ More recently it has been shown that such correlated changes occur more frequently in obligate complexes (the ones in which the two partners have to interact all the time in order to perform their biological function).⁶⁹ The hypothesis for explaining these inter-protein correlation patterns is the same as for the intra-protein ones and involves co-adaptation between the two interacting partners, in the sense that changes in one partner can be compensated by changes in the other, more probably in the regions they interact. It has been experimentally shown for some cases that compensatory changes can indeed recover the stability in complexes lost by a former mutation.⁷⁴ It is important to bear in mind that the demonstrated relationship between correlated mutations and spatial closeness (both internally and between proteins) is independent of this co-adaptation hypothesis being true.

The existence of correlated mutations between interacting proteins allows them to be used in the prediction of interacting surfaces (previous paragraph) but also in the search for the interaction partner(s) of a given protein. The idea is that interaction partners will have more correlated pairs between them and with higher correlation values. This is the basic concept behind the *in silico two-hybrid method* for locating interacting pairs of proteins⁷⁵ (Fig. 7.1F). In this method, an *interaction index* between two proteins is calculated based on the binned distributions of inter-protein and intra-protein correlation values. The interaction index between two proteins A and B is calculated as

$$C_{AB} = \sum_{i=incorr}^n \frac{P_{ABi}}{P_{Ai} + P_{Bi}} Corr_i$$

were P_{Ai} and P_{Bi} are the fractions of pairs with correlation values within bin i internal to proteins A and B respectively. P_{ABi} is the corresponding value for inter-protein pairs (pairs in which one residue belongs to protein A and the other to B). Correlation values, calculated as in Göbel *et al.*,⁷¹ are binned and the sum runs for all the bins from an initial value *incorr* up to the n th bin, which corresponds to a correlation value of 1.0. $Corr_i$ is the correlation value for bin i .

It was shown for different datasets that pairs of proteins with a high interaction index tend to be real interaction partners.⁷⁵ One advantage of this coevolution-based method with respect to the others is the possibility of obtaining information on the interaction surface concomitantly with the detection of interaction partners, because one can, from a high interaction index, go back to the actual correlated pairs of residues responsible for it. Another advantage of this method is that, due

to the residue coevolution idea behind it, it is supposed to be closer to the detection of physical interactions, in contrast to other methods which are expected to detect both physical and functional interactions. Its disadvantage is that it requires many homologous sequences of the two proteins to work, as the mirror-tree method does.

7.2.7. Other methods

There are many other evolution-based methods which use sequence or genomic features for predicting interactions. They are not extensively described here due to space limitations. The methods described so far do not involve training, that is, they do not learn from examples of known interactions and non-interactions. There is another class of methods that are trained with examples.^{28,76–79} These are sometimes termed *supervised methods*. The input for these methods is a set of characteristics (descriptors) of the proteins or protein pairs. Using a set of known protein-protein interactions, a classifier (i.e. neural net, SVM, etc.) learns to distinguish interacting from non-interacting pairs based on the values of these descriptors.

For example, Sprinzak & Margalit⁷⁸ use pairs of sequence signatures extracted from known interactions to predict new ones. Some of the methods described previously also have their supervised versions which involve training with examples.⁵⁸

7.3. Conclusion

The *ab initio* determination of interaction partners (based on basic physico-chemical principles) involves tremendous problems, maybe unsolvable ones. On the other hand, experimental techniques for the high-throughput determination of interacting pairs of proteins have many intrinsic drawbacks. One successful alternative to complement these approaches is the detection of interacting pairs of proteins by studying the landmarks left on them by the evolutionary process. Interacting proteins are intuitively expected to have particular evolutionary features (coevolution, etc.). The continuous accumulation of genomics and proteomics data makes it easier every day to trace back these evolutionary histories and hence to detect interaction partners. It has been indeed shown for some of these evolution-based methods that their accuracy increases, in general, as we use more data (i.e. the number of sequenced genomes increases).^{48,49}

The idea behind all these methods is that interacting and functionally related proteins are forced to coevolve, adapting to each other. Destabilizing or function-changing mutations in one protein could be compensated by changes in its partner (correlated mutations). A long process of such co-adaptation at the sequence level could be reflected in a similarity of evolutionary histories (similarity of phylogenetic trees), although similar evolutionary rates in the two families would also explain the observed coevolution without requiring these compensatory changes. The limit of such coevolutionary process would be to adapt not only sequence features but the existence of the proteins themselves as well, removing one partner when the

other is not present (similarity of phylogenetic profiles). Furthermore, evolution might lead to a fusion of the two proteins to increase the effective concentration of the functional complex (gene fusion), or to keep them together in the same operon to allow co-transcription (gene neighboring). These evolutionary assumptions also highlight a general limitation of these methods: they cannot be applied to heterologous interactions (i.e. antigen-antibody).

Although it is difficult to compare the different *in silico* methods for predicting protein interactions because they have different limitations in the ranges of applicability, some attempts are being made in this direction.¹⁶ The general conclusion could be that these methods have different ranges of accuracy and coverage, being the methods with highest accuracy being the ones with lowest coverage, and *vice versa*. Moreover, the type of the predicted interactions (functional, physical, neighbouring in metabolic pathways, etc.) also differs between methods in a way that is not completely clear. Since there is no method clearly better than the others, and some methods are more suitable than others for certain types of interactions, the final user has to try different ones and interpret the results in terms of what is known about the target protein. There are some repositories available online, where the user can look for the interaction partners predicted by these and other methods.^{51,80}

Establishing the complete structure of the dynamic interactome of a living cell, including the modulation of the interactions in different cellular states (temporal) and compartments (spatial), is a formidably complex problem. The characterization of the static protein interaction networks is only the first step. A combination of static information on protein interactions with information on gene expression (DNA arrays) is starting to be used to get closer to the real dynamic interactome.^{21,81}

The study of protein interaction networks is important not only from a theoretical stance but also in terms of potential practical applications, since it might enable new drugs to be developed to interrupt or modulate protein interactions instead of simply targeting a given protein's complete set of functions. Knowing the interactome may also allow a rational selection of multiple drug targets, by choosing the nodes/connections one wants to target in order to isolate or deactivate a given functional region of the interactome.

A clever combination of experimental and computational techniques for the detection of protein interactions, both with their own advantages and drawbacks, will help us to interpret the genomic information in functional terms, which is the final goal of the post-genomic era.

Acknowledgements

We thank the members of the Protein Design Group (CNB-CSIC, Madrid), specially David de Juan, and the members of the Structural Bioinformatics Group (Imperial College London), especially Prof. Michael J.E. Sternberg, for the inter-

esting discussions. This work was funded in part by the grants BIO2006-15318 and PIE 200620I240 from the Spanish Ministry for Education and Science, and the BioSapiens Network of Excellence (LSHG-CT-2003-503265).

References

1. H. Kitano, Systems biology: A brief overview, *Science*. **295**, 1662–1664, (2002).
2. P. Nurse, Systems biology: understanding cells, *Nature*. **424**, 883, (2003).
3. M. van Regenmortel, Reductionism and complexity in molecular biology. scientists now have the tools to unravel biological and overcome the limitations of reductionism, *EMBO Reports*. **5**, 1016–1020, (2004).
4. S. Fields and O. Song, A novel genetic system to detect protein-protein interactions, *Nature*. **340**, 245–246, (1989).
5. M. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, J. Schultz, J. Rick, A. Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, RuffnerH, A. Merino, M. Hudak, D. Dickson, T. Rudi, V. Ganu, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E. Querfurth, R. V, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and S.-F. G, Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*. **415**, 141–147, (2002).
6. Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskut, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. Willems, H. Sassi, P. Nielsen, K. Rasmussen, J. Andersen, L. Johansen, L. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. S?rensen, J. Matthiesen, R. Hendrickson, F. Gleeson, T. Pawson, M. Moran, D. Durocher, M. Mann, C. Hogue, D. Figeys, and M. Tyers, Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry., *Nature*. **415**(6868), 180–3, (2002).
7. J. Rain, L. Selig, H. D. Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Ghemana, A. Labigne, and P. Legrain, The protein-protein interaction map of *Helicobacter pylori*, *Nature*. **409**, 211–215, (2001).
8. G. Butland, J. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili, Interaction network containing conserved and essential protein complexes in *escherichia coli*, *Nature*. **433**, 531–537, (2005).
9. U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlauff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toks?z, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. Wanker, A human protein-protein interaction network: a resource for annotating the proteome., *Cell*. **122**(6), 957–68, (2005).
10. T. Ito, K. Tashiro, S. Muta, R. Czawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, Towards a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins., *Proc. Natl. Acad. Sci. USA*. **97**, 1143, (2000).
11. P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, V. Narayan, L. D., M. Srinivasan, P. Pochart, Q.-E. A., Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. Rothberg, A comprehensive

- analysis of protein-protein interaction networks in *saccharomyces cerevisiae*, *Nature*. **403**, 623–627, (2000).
12. S. Li, C. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. Vidalain, J. Han, A. Chesneau, T. Hao, D. Goldberg, N. Li, M. Martinez, J. Rual, P. Lamesch, L. Xu, M. Tewari, S. Wong, L. Zhang, G. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. Gabel, A. Elewa, B. Baumgartner, D. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. Mango, W. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. Gunsalus, J. Harper, M. Cusick, F. Roth, D. Hill, and M. Vidal, A map of the interactome network of the metazoan *c. elegans*., *Science*. **303**(5657), 540–3, (2004). ISSN 1095-9203.
 13. L. Giot, J. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. Stanyon, R. Finley, K. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. Shimkets, M. McKenna, J. Chant, and J. Rothberg, A protein interaction map of *drosophila melanogaster*., *Science*. **302**(5651), 1727–36, (2003).
 14. P. Aloy and R. Russell, Interrogating protein interaction networks through structural biology, *Proc. Natl. Acad. Sci. USA*. **99**, 5896–5901, (2002).
 15. P. Legrain, J. Wojcik, and J. Gauthier, Protein-protein interaction maps: a lead towards cellular functions, *Trends Genet.* **17**, 346–352, (2001).
 16. C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, Comparative assessment of large-scale data sets of protein-protein interactions., *Nature*. **417**(6887), 399–403 (May, 2002).
 17. B. Grünfelder and E. Winzeler, Treasures and traps in genome-wide data sets: case examples from yeast, *Nat. Rev. Genet.* **3**, 653–661, (2002).
 18. R. Kothapalli, S. Y. and S. Mane, and T. Loughran, Microarray results: how accurate are they?, *BMC Bioinformatics*. **3**, 22, (2002).
 19. D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, Topological structure analysis of the protein-protein interaction network in budding yeast, *Nucl. Acid Res.* **31**(9), 2443–2450, (2003).
 20. H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, Evolutionary rate in the protein interaction network., *Science*. **296**(5568), 750–2 (Apr, 2002).
 21. J. Han, N. Bertin, T. Hao, D. Goldberg, G. Berriz, L. Zhang, D. Dupuy, A. Walhout, M. Cusick, F. Roth, and M. Vidal, Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*. **430**(6995), 88–93, (2004).
 22. H. Jeong, S. Mason, A. Barabasi, and Z. Oltvai, Lethality and centrality in protein networks, *Nature*. **411**(6833), 41–42, (2001).
 23. H. Qin, H. H. S. Lu, W. B. Wu, and W.-H. Li, Evolution of the yeast protein interaction network., *Proc. Natl. Acad. Sci. USA*. **100**(22), 12820–4 (Oct, 2003).
 24. S. Wuchty and P. F. Stadler, Centers of complex networks., *J Theor Biol.* **223**(1), 45–53 (Jul, 2003).
 25. E. Yeger-Lotem and H. Margalit, Detection of regulatory circuits by integrating the cellular networks of protein-protein interactions and transcription regulation, *Nucl. Acid Res.* **31**, 6053–6061, (2003).
 26. M. Gomez, R. Alonso-Allende, F. Pazos, O. Grana, D. Juan, and A. Valencia. Accessible protein interaction data for network modeling. structure of the information

- and available repositories. In ed. C. Priami, *Transactions on Computational Systems Biology I: Subseries of Lecture Notes in Computer Science*, pp. 1–13. Springer, (2005).
27. M. Lappe and L. Holm, Unraveling protein interaction networks with near-optimal efficiency., *Nat. Biotechnol.* **22**(1), 98–103 (2004).
 28. R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M. Gerstein, A bayesian network approach for predicting protein-protein interactions from genomic data, *Science*. **302**, 449–453, (2003).
 29. A. Archakov, V. Govorun, A. Dudanov, Y. Ivanov, A. Veselovsky, P. Lewi, and P. Janssen, Protein-protein interactions as a target for drugs in proteomics, *Proteomics*. **3**, 380–391, (2003).
 30. R. Russell, F. Alber, P. Aloy, F. Davis, M. Pichaud, M. Topf, and A. Sali, A structural perspective on protein-protein interactions, *Curr. Opin. Struct. Biol.* **14**, 313–324, (2004).
 31. L. Salwinski and D. Eisenberg, Computational methods of analysis of proteinprotein interactions, *Curr. Opin. Struct. Biol.* **13**, 377–382, (2003).
 32. A. Szilágyi, V. Grimm, A. Arakaki, and J. Skolnick, Prediction of physical protein-protein interactions, *Phys. Biol.* **2**, S1–S16, (2005).
 33. J. Janin and B. Seraphin, Genome-wide studies of protein-protein interaction, *Curr. Opin. Struct. Biol.* **13**, 383–388, (2003).
 34. G. Smith and M. Sternberg, Prediction of protein-protein interactions by docking methods, *Curr. Opin. Struct. Biol.* **12**, 28–35, (2002).
 35. P. Uetz and R. Finley, From protein networks to biological systems, *FEBS Lett.* **579**, 1821–1827, (2005).
 36. R. Hoffmann and A. Valencia, Protein interaction: same network, different hubs., *Trends Genet.* **19**(12), 681–3 (Dec, 2003).
 37. J. Tamames, G. Casari, C. Ouzounis, and A. Valencia, Conserved clusters of functionally related genes in two bacterial genomes, *J. Mol. Biol.* **44**, 66–73, (1997).
 38. T. Dandekar, B. Snel, M. Huynen, and P. Bork, Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem. Sci.* **23**, 324–328, (1998).
 39. R. Overbeek, M. Fonstein, M. D’Souza, G. Pusch, and N. Maltsev, Use of contiguity on the chromosome to predict functional coupling, *In Silico Biol.* **1**, 93–108, (1999).
 40. A. Enright, I. Iliopoulos, N. Kyripides, and C. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, *Nature*. **402**, 86–90, (1999).
 41. E. Marcotte, M. Pelligrini, M. Thompson, T. Yeates, and D. Eisenberg, A combined algorithm for genome-wide prediction of protein function, *Nature*. **402**, 83–86, (1999).
 42. S. Tsoka and C. Ouzounis, Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion, *Nat. Genetics*. **26**(141-142), (2000).
 43. T. Gaasterland and M. Ragan, Microbial genescapes: phyletic and functional patterns of orf distribution among prokaryotes, *Microb. Comp. Genomics*. **3**, 199–217, (1998).
 44. M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg, and T. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles., *Proc. Natl. Acad. Sci U S A*. **96**(8), 4285–8, (1999).
 45. S. Altshul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, Gapped blast and psi-blast: a new generation of protein database search programs, *Nucl. Acid Res.* **25**, 3389–3402, (1997).
 46. S. Date and E. Marcotte, Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages., *Nat. Biotechnol.* **21**(9), 1055–62, (2003).
 47. C. Shannon and W. Weaver, *The Mathematical Theory of Communication*. (University of Illinois Press, 1962).
 48. J. Sun, J. Xu, Z. Liu, Q. Liu, A. Zhao, T. Shi, and Y. Li, Refined phylogenetic

- profiles method for predicting protein-protein interactions, *Bioinformatics*. **21**, 3409–3415, (2005).
49. Y. Zheng, R. Roberts, and S. Kasif, Genomic functional annotation using coevolution profiles of gene clusters, *Genome Biology*. **3**, 61–69, (2002).
 50. E. Morett, J. Korb, E. Rajan, G. Saab-Rincon, L. Olvera, S. Schmidt, B. Snel, and P. Bork, Systematic discovery of analogous enzymes in thiamin biosynthesis, *Nat. Biotechnol.* **21**, 790–795, (2003).
 51. P. Bowers, S. Cokus, D. Eisenberg, and T. Yeates, Use of logic relationships to decipher protein network organization., *Science*. **306**(5705), 2246–9, (2004).
 52. K. Fryxell, The coevolution of gene family trees, *Trends Genet.* **12**, 364–369, (1996).
 53. S. Pages, A. Belaich, J. Belaich, E. Morag, R. Lamed, Y. Shoham, and E. Bayer, Species-specificity of the cohesin-dockerin interaction between *clostridium thermocellum* and *clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain, *Proteins*. **29**, 517–527, (1997).
 54. H. Fraser, A. Hirsh, D. Wall, and M. Eisen, Coevolution of gene expression among interacting proteins, *Proc. Natl. Acad. Sciences USA*. **101**, 9033–9038, (2004).
 55. C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen, Co-evolution of proteins with their interaction partners., *J. Mol. Biol.* **299**(2), 283–293 (Jun, 2000).
 56. F. Pazos and A. Valencia, Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Engineering*. **14**, 609–614, (2001).
 57. F. Pazos, J. Ranea, D. Juan, and M. Sternberg, Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome., *J. Mol. Biol.* **352**(4), 1002–15, (2005).
 58. R. Craig and L. Liao, Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices, *BMC Bioinformatics*. **8**, 6, (2007).
 59. J. Gertz, G. Elford, A. Shustrova, M. Weisinger, M. Pellegrini, S. Cokus, and B. Rothschild, Inferring protein interactions from phylogenetic distance matrices, *Bioinformatics*. **19**, 2039–2045, (2003).
 60. J. Izarzugaza, D. Juan, C. Pons, J. Ranea, A. Valencia, and F. Pazos, Tsema: interactive prediction of protein pairings between interacting families, *Nucl. Acid Res.* **34**, W315–319, (2006).
 61. R. Jothi, M. Kann, and T. Przytycka, Predicting protein-protein interaction by searching evolutionary tree automorphism space, *Bioinformatics*. **21**, i241–i250, (2005).
 62. D. Juan, F. Pazos, and A. Valencia, High-confidence prediction of global interactomes based on genome-wide coevolutionary networks, *Proc. Natl. Acad. Sci. U S A*. **105**, 934–939, (2008).
 63. M. Kann, R. Jothi, P. Cherukuri, and T. Przytycka, Predicting protein domain interactions from coevolution of conserved regions, *Proteins*. **67**, 811–820, (2007).
 64. W. Kim, D. Bolser, and J. Park, Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (psimap), *Bioinformatics*. **20**, 1138–1150, (2004).
 65. A. Ramani and E. Marcotte, Exploiting the co-evolution of interacting proteins to discover interaction specificity, *J. Mol. Biol.* **327**, 273–284, (2003).
 66. T. Sato, Y. Yamanishi, K. Horimoto, H. Toh, and M. Kanehisa, Prediction of protein-protein interactions from phylogenetic trees using partial correlation coefficient, *Genome Informatics*. **14**, 496–497, (2003).
 67. T. Sato, Y. Yamanishi, M. Kanehisa, and H. Toh, The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships, *Bioinformatics*. **21**, 3482–3489, (2005).

68. S. Tan, Z. Zhang, and S. Ng, Advice: Automated detection and validation of interaction by co-evolution, *Nucl. Acid. Res.* **32**, W69–W72, (2004).
69. J. Mintseris and Z. P. Weng, Structure, function, and evolution of transient and obligate protein-protein interactions, *Proc. Natl. Acad. Sci. U S A.* **102**(31), 10930–10935 (Aug., 2005).
70. H. Jothi, P. Cherukuri, A. Tasneem, and T. Przytycka, Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions, *J. Mol. Biol.* **362**, 861–875, (2006).
71. U. Göbel, C. Sander, R. Schneider, and A. Valencia, Correlated mutations and residue contacts in proteins, *Proteins.* **18**, 309–317, (1994).
72. O. Olmea and A. Valencia, Improving contact predictions by the combination of correlated mutations and other sources of sequence information, *Fold. Des.* **2**, S25–S32, (1997).
73. F. Pazos, M. HelmerCitterich, G. Ausiello, and A. Valencia, Correlated mutations contain information about protein-protein interaction, *J. Mol. Biol.* **271**(4), 511–523 (Aug., 1997).
74. M. Mateu and A. Fersht, Mutually compensatory mutations during evolution of the tetramerization domain of tumor suppressor p53 lead to impaired hetero-oligomerization, *Proc. Natl. Acad. Sci. USA.* **96**, 3595–3599, (1999).
75. F. Pazos and A. Valencia, In silico two-hybrid system for the selection of physically interacting protein pairs, *Proteins-Structure Function And Genetics.* **47**(2), 219–227 (May, 2002).
76. A. Ben-Hur and W. Noble, Kernel methods for predicting protein-protein interactions, *Bioinformatics.* **21**, i38–46, (2005).
77. X. Chen and M. Liu, Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics.* **21**, 4394–4400, (2005).
78. E. Sprinzak and H. Margalit, Correlated sequence-signatures as markers of protein-protein interactions, *J. Mol. Biol.* **311**, 681–692, (2001).
79. Y. Yamanishi, J. Vert, and M. Kanehisa, Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics.* **20**, I363–I370, (2004).
80. C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, String: a database of predicted functional associations between proteins, *Nucl. Acid Res.* **31**, 258–261, (2003).
81. U. de Lichtenberg, L. Jensen, S. Brunak, and P. Bork, Dynamic complex formation during the yeast cell cycle, *Science.* **307**, 724–727, (2005).

This page intentionally left blank

Chapter 8

Statistical Null Models for Biological Network Analysis

WILLIAM P. KELLY, THOMAS THORNE AND MICHAEL P.H. STUMPF

*Centre for Bioinformatics, Division of Molecular Biosciences,
Imperial College London*

*william.kelly04@imperial.ac.uk, thomas.thorne@imperial.ac.uk,
m.stumpf@imperial.ac.uk*

Statistical ensembles of random graphs serve as null models in the statistical analysis of real complex networks. They encapsulate what are believed to be the generic properties of networks and describe the expected behaviour against which observed network data can be compared. Here we review the basic statistical physics underlying statistical ensembles of networks and show how we can exploit their properties. We also show how the simple statistical ensembles that have been used to describe networks can be improved by conditioning the ensembles on other available data. We show that such conditional ensembles provide biologically more realistic network null models which can be used for more detailed functional and evolutionary analyses.

8.1. Introduction

Molecular interaction and regulatory networks have taken a central role in bioinformatics and the fledgling field of systems biology: they provide concise and comprehensive descriptions of the molecular machinery underlying biological processes, are amenable to mathematical and statistical analysis and modelling, and can visualize complex relationships among the constituents of cellular systems. For these reasons they can offer a convenient link between mathematical analysis and biological understanding.

In this chapter we present a statistical perspective on how to analyze biological network data. In particular we will address fundamental yet simple questions such as:

- how similar are the properties of interacting proteins?
- is the available protein-interaction data a fair representation of the overall interaction data?

Such questions are closely related to the bread-and-butter problems of conventional statistics, but the network introduces dependencies among the nodes in the network

which may render many of the standard statistical tests (such as basic hypothesis testing) useless or inadequate.^{1,2} There has, for example, been considerable debate as to whether interacting proteins coevolve.^{3–5} This is a question both of fundamental evolutionary interest as well as practical importance; if interacting proteins do evolve in a concerted manner then this would potentially help in determining protein-protein interactions from phylogenetic information. But its answer depends, as we will show below, on how we choose to include the network into the analysis. The dependencies that exist between nodes in the network affect analyses in a similar manner as is the case for data on trees, e.g. in phylogenetics.⁶ But while efficient algorithms exist for dealing with tree data, reticulations and loops in the network which give rise to many different routes between pairs of nodes in a network introduce considerable computational problems for the mathematical and statistical analysis.

Our understanding of protein interaction networks has grown rapidly over the past 10 years but we feel it is regrettable that so many results from the early days, which have since been shown to be incorrect, are still floated and accepted in parts of the community. As our knowledge of these networks has increased, so has our knowledge of other forms of biological data. In order to yield truly meaningful results we have to combine and fuse these different types of information. Here we will review recent developments in this area from a statistical perspective.

8.1.1. *Protein interaction networks*

Protein interaction networks, at least in their current guise, provide a static representation of the physical interactions in biological organisms. Whereas physical protein-protein interactions will change over time and in response to environmental, developmental and physiological cues, present network representations fail to acknowledge that. Rather we view a PIN as the union of a set of nodes, $\mathcal{N} = \{n_1, n_2, \dots, n_N\}$, corresponding to the N proteins in an organism, and the set of PPIs, $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$, where $e_k = e_{ij}$ if $n_i, n_j \in \mathcal{N}$ and an interaction between proteins n_i and n_j has been reported.

Data comes in two guises: some experimental techniques detect evidence for direct pairwise physical interactions between proteins or protein domains. Other techniques, based on mass-spectrometric assays, identify sets of proteins which interact together, without necessarily being able to disentangle them into pairwise interactions. Several databases^{7,8} contain protein interaction data, with a notable bias in favour of model organisms and, more recently, humans. For non-model organisms data is generally restricted to *in silico* inferences of interactions, typically exploiting homology arguments.

8.1.2. Statistical analysis of network data

Present protein interaction data sets are limited to static representations of *in vitro* interactions, but recent progress in mapping interactions under more realistic conditions promises to change our understanding of interactions considerably.⁹ Because of experimental limitations and challenges the data is, however, of a somewhat preliminary nature. But this and the fact that interactome data is highly incomplete and plagued by considerable false positive and false negative rates, have been ignored in the vast majority of analyses.^{10,11} Generally, such aspects of the data ought to be included into the analysis as both the incomplete nature and the unreliability of PPI information can have profound influence on the insights that can be gained from such data.

Statistical tools are being developed to clean up PPI data, to predict PPI data using a range of statistical learning approaches and to evaluate the properties of PINs and their organization in light of evolutionary mechanisms or available additional biological data. All of these have been studied extensively in the literature (including chapters in this book). Here we take a slightly more detached perspective and discuss how we can construct suitable null models for the statistical analysis of biological network data. Null models play a central part in frequentist statistics, in particular in the context of hypothesis testing. A null hypothesis is a plausible probability model or process which could have generated the observed data. While we are never able in frequentist statistics to accept the null model, we may be able to reject it as implausible in light of the available data.¹²

More generally, and going beyond the limitations imposed by frequentist hypothesis testing, we can also use different models of network evolution or organization,^{13–15} compare them in light of the available data, and either choose the best model or average over predictions from all models (weighted by the statistical evidence in their favour). In all cases we can and should employ the notion of network ensembles or probability spaces over graphs. We will introduce these concepts in the next section in a semi-formal manner before employing them in the context of the *S. cerevisiae* PIN. There we shall study the issue of coevolution of interacting proteins from different perspectives before briefly considering how the network data has been collected over time.

8.2. Network Ensembles

The notion of a statistical ensemble^{16–19} is closely aligned to statistical analysis and, in particular, natural from a Bayesian point of view. Very loosely speaking, we consider each network as belonging to a set of networks with similar (or identical) properties. More formally, an ensemble is the set of all possible microscopic states a system can take under a certain constraint. By considering a given instance of a network as part of an ensemble of networks we can compare systematically its properties to those of the networks in the ensemble in general. For a given

ensemble of systems X we assume that the probability of a particular ensemble member $x \in X$ is given by $\Pr(x)$, whence the ensemble average of some property S of X is given by

$$\langle S \rangle = \frac{1}{Z} \sum_{x \in X} S(x) \Pr(x)$$

where $Z = \sum_{x \in X} \Pr(x)$ is generally known as the partition function.

The ensemble thus serves as a useful null model for our analysis and further hypothesis testing. Below we will provide a brief and self-contained review of ensembles in statistical physics before defining a general and mathematically stripped-down version of a class of random network ensembles which we believe is particularly suited to network analysis. We will conclude this section with a brief outline of how to go beyond simple network ensembles, a thread which is picked up again in the following sections.

8.2.1. *Ensembles in statistical physics*

Whereas we can easily describe the behaviour of a single particle (at least in classical physics) in terms of fundamental equations of motion, this perspective breaks down as we consider larger and larger number of particles.¹⁸ For N particles in three-dimensional space we require $6N$ variables to describe their microscopic states (for each particle we need the 3 coordinates and the moments in the three directions). Following the pioneering work of Ludwig Boltzmann who considered, very much against contemporary fashion, the statistical properties of ensembles of identical particles, theoretical physics has made enormous progress by likening macroscopic phenomena to a statistical treatment of microscopic dynamics.

We define ensembles in terms of features or properties that are conserved among all members of the ensemble. Three types of ensemble are generally being considered and we adopt the physics terminology.

Micro-canonical ensemble: In conventional physics the total energy and number of particles are conserved. A micro-canonical network ensemble is defined by an sequence of integers, $\{n_0, n_1, \dots, n_t\}$ with $0 < t \leq N$, where n_k is the number of nodes in the network that have k incident edges such that

$$\sum_{k=0}^t n_k = N \quad \text{and} \quad \sum_{k=0}^t k n_k = 2M.$$

Each network $\mathcal{N}$ which fulfils these conditions is given equal statistical probability, $\Pr(\mathcal{N}) = \text{const.}$

Canonical ensemble: Total energy may thermally fluctuate subject to a constant temperature and fixed number of particles. In a network context, networks belonging to the canonical ensembles have a fixed number of edges and are characterized by a probability distribution for the degree sequence; now the probability

of a node having degree k is given by $p(k)$. In the thermodynamic limit (i.e. as $N \rightarrow \infty$) the definitions for micro-canonical and canonical ensembles used here become equivalent.

Grand canonical ensemble: In statistical physics the temperature and the chemical potential (the expected number of particles) are fixed. In a network context this corresponds to the case where we only specify the probability distribution for the degree sequence $p(k)$; thus the number of edges in the network, M , is now allowed to vary.

For example, classical Erdős-Rényi random graphs^{20,21} where M edges are randomly distributed among N nodes form a canonical ensemble, whereas the related classical random graph model originally conceived by Gilbert,²² where each pair of nodes is connected with constant probability p forms a grand canonical ensemble of networks.

There are different ways of defining these network ensembles but the current approach is particularly useful and we will discuss networks in this framework. Equivalently we could speak of probability spaces over networks instead of ensembles. We note that throughout this chapter we choose to ignore potential issues arising from multiple interactions among pairs of nodes or self-interactions of a node with itself. Biologically, however, the latter in particular will frequently have to be considered.

8.2.2. *Bender-Canfield (BC) networks*

The classical example of a micro-canonical network ensemble is due to Bender and Canfield²³ who considered properties of networks which are defined in terms of a given degree sequence, $n(k)$. We will call this type of graph a Bender-Canfield or BC graph (see Fig. 8.1). We can think of the BC ensemble as a set of N nodes where $n(k)$ is the number of nodes with k stubs which are wired up randomly. In practice we pick without replacement pairs of stubs and connect them by an edge until all edges have been distributed and no free stubs remain.

We will consider BC ensembles in the thermodynamic limit ($N \rightarrow \infty$); here, because the different ensembles become equivalent, the BC ensemble properties are of course the same as those of an ensemble where only the degree probability distribution (but not the sequence itself) is fixed. We will therefore take the notational liberty of considering the case of fixed degree distribution $\text{Pr}(k)$ rather than merely a fixed degree sequence $n(k)$.

BC graphs have gained popularity because they allow some analytical insight into the global characteristics of networks, in particular as $N \rightarrow \infty$. The most prominent example of such analytical results is the *Molloy-Reed criterion*^{24,25} which states that as $N \rightarrow \infty$ a network will have a giant connected component if and only if the number of next nearest neighbours is larger than the number of nearest neighbours (provided both are finite numbers); here the giant connected component is a set of nodes that can all be reached from one another by traversing along edges

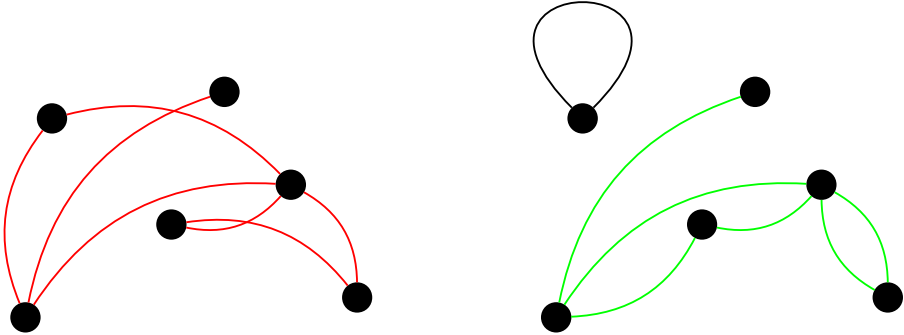


Fig. 8.1. Two networks with the same degree sequence which belong to the BC ensemble characterized by the degree sequence, $k \in \{1, 2, 2, 2, 3, 4\}$. In the general ensemble we do not disregard networks with multiple edges and/or loops.

in the network connecting these nodes.

Generally, the bulk of statistical analyses compare the observed networks with random networks drawn from a BC ensemble. This is understandable given the ease with which these confidence intervals are being generated. However, this perspective has apparently mostly been adopted without any further consideration of the concomitant limitations. This is particularly the case for several of the earlier analyses on PIN data which, despite limitations in the data available to them and a certain lack of statistical rigour in some cases, continue to be cited in the literature uncritically.

8.2.3. *Beyond BC networks*

The ensemble of BC networks has many attractive features; most importantly it allows for comprehensive analytical analyses as in the limit where $N \rightarrow \infty$, the effects of loops and closed paths can be ignored.²⁶ The graphs drawn from a BC ensemble do, however, ignore correlations observed in real networks. These correlations can be due to biological organization or be induced by the evolutionary process which gave rise to the network. These two factors are of course intimately linked but can be (artificially) separated for the sake of easing the analysis.

For computational convenience we typically treat these two aspects separately. Below we will show how ensembles of networks can be generated that condition on additional biological knowledge about the makeup of biological organisms. We may for example want to condition our rewired networks not only on the degree distribution, but also on the clustering coefficient²⁷ or degree-degree distribution $\Pr(k, k')$, the probability that a node with degree k interacts with a node with degree k' .

The most important deviation from BC networks probably originate from the process by which the networks have evolved.¹⁵ Different evolutionary processes give rise to different levels of correlations among interacting nodes. For example, a

process involving duplication of nodes and all their edges with subsequent removal or rewiring of existing edges or addition of new edges will tend to give rise to networks with high clustering coefficients.

Most network growth models are modelled as Markov chains and the degree distribution can generally be calculated from a suitable master equation²⁸

$$\Pr(k, t) = \sum_{i=0}^{N_t} (M_{i,k} \Pr(i, t-1) - M_{k,i} \Pr(k, t-1)), \quad (8.1)$$

where $\Pr(k, t)$ is the probability of a node having degree k at time t . If we add one node at each time-point then the number of nodes at time t is $N_t = t$; $M_{i,k}$ is the probability of going from degree i to degree k . To each such growth model we will thus be able to assign a corresponding BC ensemble given the degree sequence which can be obtained from the master equation.

So far, all studies of which we are aware have assumed a stationary Markov process. From evolutionary biology, however, we know that the manner in which real networks have grown or in which organismic complexity has shifted over time is (i) highly contingent, (ii) diverse, and (iii) not gradual but characterized by a sequence of major evolutionary events. Such events include well documented whole genome duplications and presumably a host of smaller events such as duplication or deletion of chromosomal segments.

To capture the correlations, etc. in growing networks we either have to use a model-based approach where we generate networks using one or more hypothetical growth mechanisms,^{14,29} or we have to start with a BC ensemble and condition the network on the additional data by selectively rewiring edges. Below we illustrate an approach that goes beyond the simple rewiring by developing a Markov chain which explicitly conditions on available functional data.

8.3. Generating Confidence Intervals on Networks

Given a set of nodes, $\mathcal{V}^*$, and the reported interactions among these nodes, $\mathcal{E}^*$, we want to determine if some nodal properties, $c_i \forall i \in \mathcal{V}^*$, are for instance more similar among interacting nodes than among non-interacting nodes. Here the c_i could, for example, be the evolutionary rate of a protein, its phylogeny across a panel of related species, the expression level, or any other annotation of the protein.

We will use the concept of BC graphs introduced above in order to formalize the vague notion of similarity among nodes in a network. We always assume that the structure of the observed network $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ is given in terms of the adjacency matrix $A^* = (a)_{ij}$ with $a_{ij} = 1$ if nodes v_i and v_j are connected by an edge and 0 otherwise; i.e. we assume binary interactions and thus have no qualitative or temporal data on the edges. In each case we calculate some statistic of the observed network (such as the Pearson correlation of the expression levels of interacting proteins) and for a range of networks generated under one of the Null models below.

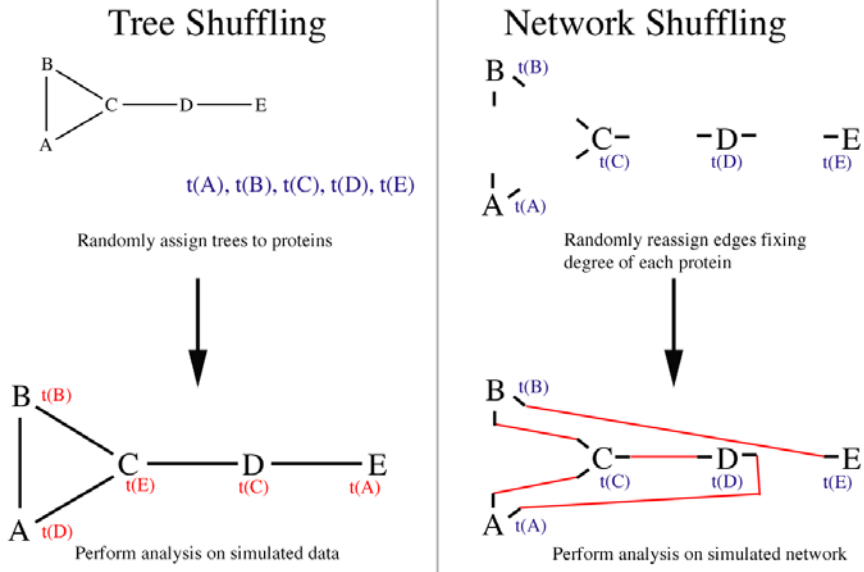


Fig. 8.2. Descriptions of how the random networks are generated through use of the *Network Shuffle* and *Tree Shuffle* null models.

8.3.1. *Random permutation of node properties* — NodeShuffle

In the first instance we may choose to keep the adjacency matrix fixed, i.e. for all q networks in the (finite, $q \leq N!$) ensemble we have

$$A_s = A^*, \quad 1 \leq s \leq q.$$

Rather we randomly permute the c_i , $1 \leq i \leq N$. This approach keeps the network fixed, including all local neighbourhoods and correlations among degrees (including the clustering coefficient) but breaks up the link between the properties under consideration and the degree of the node.

NodeShuffle (see Fig. 8.2) provides a statistical null model for the organization of functional characteristics of network nodes which can be used to test for a link between the degree of a node i , k_i , and its property (or properties), c_i .

When we consider only pairwise correlations or measures of pairwise similarity then **NodeShuffle** reduces in fact to a general, unstructured permutation test, where the set of characteristics, c_i , is shuffled randomly and pairs of entries are compared. Only when we consider network features such as cliques, closed triangles etc., does it become a truly network aware statistical tool.

8.3.2. *Random rewiring of networks*

The alternative to permuting the assignment of characteristics to nodes is to permute or randomize the structure of the network itself. There are three options of

doing this: (i) we can randomize the M edges among the N nodes, (ii) we can randomly rewire the edges keeping the node of each degree fixed, or (iii) we can rewire the nodes such that their degree is fixed while also maintaining other characteristics of the network (such as community structure).

The first option, which assumes that the correct Null model for the network is a classical or Erdős Rényi random graph, is not relevant in a biological context where the node degree distribution is generally far from Poisson. We therefore focus here on the remaining two. We will consider all three approaches again at the end of this section.

8.3.2.1. Random rewiring of networks — NetShuffle

If we want to keep the link between node degree and characteristics fixed, as should be done, if there is reason to believe that the degree is a confounding variable for that characteristic, then we need to consider different null models. The most commonly used approach is to implicitly consider the observed network in the context of its BC ensemble (see Fig. 8.2). That is, we compare the statistics observed in our given network against the statistics obtained in networks that are characterized by the same degree distribution and the same mapping of characteristics c_i onto nodes v_i .

To this end all we have to do is follow a procedure that generates networks that belonging to the same BC ensemble as the true network. And random rewiring of edges, keeping the degree of each node fixed, achieves just this.

8.3.2.2. Conditional rewiring of networks — G0cardShuffle

In most biological contexts (or in real networks in general) there is substantial additional structure in the network: proteins tend to interact predominantly with proteins that are localized in the same cellular component, involved in the same biological process or have the same or similar biological function. For many organisms, in particular *S. cerevisiae*, such functional annotations are accessible in gene ontologies (GO). Clearly, the random rewiring discussed above fails to take this into account. Failing to account for this available information may, however, bias our analysis.³⁰

Extending the notation used thus far we now denote by γ the set of annotations (e.g. different protein functions), and let $\gamma(i)$ be the annotation of node i . For $x, y \in \gamma$ we define ν_{xy} to be the number of edges that connect a node with annotation x to a node with annotation y . Then the probability of picking a random stub on a node with annotation x that has an edge attached leading to a node with annotation y (we say that the edge is of type (x, y)) is given by

$$\omega_{xy} = \frac{\nu_{xy}}{2M} \quad \text{for } x \neq y \quad (8.2)$$

and

$$\omega_{xx} = \frac{\nu_{xx}}{M} \quad \text{otherwise.} \quad (8.3)$$

This definition means that the probabilities are properly normalized, i.e. $\sum \omega_{xy} = 1$, where the sum runs over all pairs of indices $1 \leq x, y \leq |\gamma^k|$. If $\#x$ denotes the number of x , then normalization follows from the relationship

$$\frac{1}{M} \left(\frac{1}{2} \# \text{ edges of type}(x, y) + \frac{1}{2} \# \text{ edges of type}(y, x) + \# \text{ edges of type}(x, x) \right) \\ = \sum_{x \neq y} \omega_{xy} + \sum_x \omega_{xx} = 1 \quad (8.4)$$

because the first sum on the RHS of Eqn. (8.4) runs over all ordered pairs of distinct annotations x and y . We approximate the likelihood of a given network $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ (where $\mathcal{V}$ and $\mathcal{E}$ denote the sets of nodes and edges, respectively) as the product of the probability of edges conditional on the annotations of the nodes incident on the edge. The probability of an edge, $e(i, j)$ between two nodes with annotations $\gamma(i)$ and $\gamma(j)$ is given by $\omega_e := \omega_{\gamma(i)\gamma(j)}$, whence we approximate $\Pr(\mathcal{N}) \approx \Pr(\mathcal{E})$ and we thus have for our likelihood of the network

$$\mathcal{L}(\mathcal{N}) = \Pr(\omega|\mathcal{N}) \approx \prod_{e \in \mathcal{E}} \omega_e. \quad (8.5)$$

Given a configuration, $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ we propose a novel configuration $\mathcal{N}' = (\mathcal{V}, \mathcal{E}')$ (the set of nodes does not change, hence $\mathcal{N}' = \mathcal{N}$) by choosing two edges, $e, f \in \mathcal{E}$, at random. We consider the ordered tuple of their annotations (u, v) and (x, y) , respectively and propose new edges by swapping the edges between the nodes (see Fig. 8.3) to obtain edges e' and f' which will be of type (x, v) and (u, y) , respectively. The likelihood ratio is thus

$$\frac{\mathcal{L}(\mathcal{N}')}{\mathcal{L}(\mathcal{N})} = \frac{\prod_{e \in \mathcal{E}'} \omega_e}{\prod_{e \in \mathcal{E}} \omega_e} = \frac{\omega_{e'} \omega_{f'}}{\omega_e \omega_f}, \quad (8.6)$$

as all other edges in $\mathcal{E}$ and $\mathcal{E}'$ remain unaffected by the proposed change.

We start from a random rewiring of the network which only conserves the degree of each node. The rewiring algorithm is based on Markov Chain Monte Carlo (MCMC) approach using Metropolis sampling,^{31,32} and begins with a randomly rewired network with the desired degree sequence. A pair of edges $e = (i, j), f = (r, s)$ is chosen randomly and the incident nodes are found to have annotations $\gamma(i), \gamma(j)$ and $\gamma(r), \gamma(s)$, respectively, in the κ different categories. Then the probability of the original and the rewired networks differ only by the weights of the involved edges. The probability of accepting the new configuration $e' = (i, s), f' = (j, r)$ is thus given by the Metropolis criterion

$$p = h(\mathcal{N}, \mathcal{N}') = \min \left(1, \frac{\mathcal{L}(\mathcal{N}')}{\mathcal{L}(\mathcal{N})} \right) = \min \left(1, \frac{\omega'_{e'} \omega'_{f'}}{\omega_e \omega_f} \right). \quad (8.7)$$

The configuration remains unchanged with probability $1 - p$, whence a new configuration change will be proposed.

It is easy to see that the ensemble of networks which condition on the observed edge weights, ω , form the stationary distribution of the Markov chain thus constructed. To show this we let $r(\mathcal{N} \rightarrow \mathcal{N}')$ be the transition mechanism of the chain,

$$r(\mathcal{N} \rightarrow \mathcal{N}') = q(\mathcal{N} \rightarrow \mathcal{N}') \times h(\mathcal{N}, \mathcal{N}'), \quad (8.8)$$

where $q(\mathcal{N} \rightarrow \mathcal{N}')$ is the probability of going from network $\mathcal{N}$ to $\mathcal{N}'$. Here this step will always involve swapping of two edges. These, however, are chosen uniformly at random and therefore

$$q(\mathcal{N}' \rightarrow \mathcal{N}) = q(\mathcal{N} \rightarrow \mathcal{N}'). \quad (8.9)$$

With this it is trivial to show that the detailed balance³³ is fulfilled, i.e.

$$\begin{aligned} L(\mathcal{N})r(\mathcal{N} \rightarrow \mathcal{N}') &= L(\mathcal{N})q(\mathcal{N} \rightarrow \mathcal{N}')h(\mathcal{N}, \mathcal{N}') \\ &= L(\mathcal{N})q(\mathcal{N} \rightarrow \mathcal{N}') \min\left(1, \frac{\mathcal{L}(\mathcal{N}')}{\mathcal{L}(\mathcal{N})}\right) \\ &= q(\mathcal{N} \rightarrow \mathcal{N}') \min(\mathcal{L}(\mathcal{N}), \mathcal{L}(\mathcal{N}')) \\ &= L(\mathcal{N}')q(\mathcal{N}' \rightarrow \mathcal{N})h(\mathcal{N}', \mathcal{N}) = L(\mathcal{N}')r(\mathcal{N}' \rightarrow \mathcal{N}). \end{aligned} \quad (8.10)$$

Thus `G0cardShuffle`, because of the general properties of MCMC,^{32,33} will result in a Markov chain which has as its stationary distribution the ensemble of networks (defined by $\Pr(\omega|\mathcal{N})$) which condition on the degree sequence (by virtue of fixing the degree of each node) and on the weight matrix ω (by construction of the chain).

As in all MCMC approaches it is important to run the algorithm for a sufficiently long period to remove dependence on the initial configuration and to reach the stationary distribution of the Markov process (the *burn-in period*). After that the chain produces highly correlated configurations so configurations are sampled only after a sufficiently large number of steps in the chain (this is referred to as the *thinning-out interval*).^{33,34} Choice of the length of burn-in and thinning-out intervals require experimentation and/or fine-tuning. In `G0cardShuffle` the default parameter for the burn-in period is $100 \times M$ steps, while the thinning-out interval has a length of $10 \times M$ steps.

8.4. Analysis of Coevolution of Yeast Proteins

In the absence of population genetic data, comparisons between species in which extensive PIN data are available and (preferably closely related) other species have been used to identify potential links between the role or position of proteins in the PIN and their evolutionary properties. Relative sequence conservation or other

measures of the evolutionary rate have been used to evaluate the role of protein-protein interactions (PPI) in modulating the evolutionary properties of proteins. While initial studies³⁵ suggested that the evolutionary rate of a protein decreases as the number of its PPIs increases (as always in evolutionary analyses, such trends are associated with high variance), more extensive later studies have suggested that other factors such as the expression level or protein abundance show much stronger association with evolutionary rate than a protein's degree.^{4,36,37}

While there appears to be little evidence for the evolutionary rate to correlate strongly with the number of interactions, several studies have reported a higher than expected correlation between the evolutionary rate of interacting proteins. Generally, chemokines and their corresponding receptors have been demonstrated to show evidence for correlated evolutionary behaviour which is reflected by the similarity of their respective molecular phylogenetic trees.³⁸ In the case of *tgf β* ligands and their receptors,³⁹ the topological similarities between the protein families' phylogenies have been used successfully to predict PPIs. Additional evidence comes from studies of the *S. cerevisiae* PIN where it has been shown that duplicated genes tend to preserve the same interactions for millions of years rather than hundreds of million years.⁴⁰

The reports of such coevolution have given rise to a range of tools for the prediction of PPIs which use evolutionary arguments.⁴¹ Protein phylogenetic profiles,⁴² distance matrices^{43–45} and other measures of coevolution between proteins^{3,38,39,46} have been used to predict interactions between proteins. Phylogenetic profiling⁴² emerged as whole genome sequences became widely available. These profiles are n -bit strings for each protein where each bit indicates the existence (if the bit is in state 1) or absence (state 0) of a protein homologue in a related species (see Fig. 8.4). Such profiles have been used to infer the complexes or pathway in which an unknown protein participates, or help with predicting protein function.

In Fig. 8.3 we evaluate the hypotheses that (i) the phylogenies of interacting proteins are more similar than would be expected by chance and (ii) that the rates of interacting proteins are correlated. *A priori* we would expect some concordance among the evolutionary properties of interacting proteins. Gene trees, for example, should tend to follow the (generally accepted) species tree.^{47,48} Whether or not the phylogenetic trees, especially their topology, show evidence for co-evolution between interacting proteins more than would be expected by chance has not been tested on a global level. Here we present such a statistical analysis for the available protein-protein interaction network data in *S. cerevisiae*. As it turns out we fail to find any significant evidence for phylogenies of interacting proteins to show increased levels of similarity even under simple null models. We then investigate whether the evolutionary rates of interacting proteins show evidence for higher than expected similarities and find this to be the case under the assumption of a BC ensemble null model but not when we apply the `GOCARDSHUFFLE` null model.

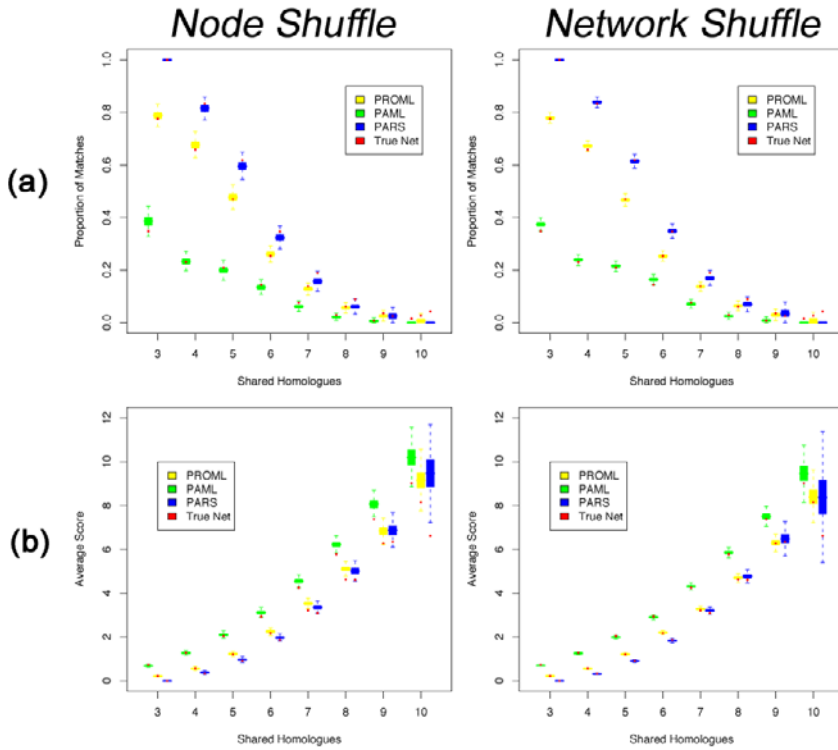


Fig. 8.3. Four boxplots show the results for the two null models; *Tree Shuffle* and *Network Shuffle* for the phylogenetic study. (a) details the proportion of matching topologies over the tree construction methods for comparisons sharing a fixed number of homologous proteins. (b) shows the average similarity score between interacting proteins over the range of shared homologues.

8.4.1. Phylogenetic analysis

Analysis was performed on different interaction datasets and using a range of phylogeny inference approaches: PROML and PARS from the Phylip 3.6⁴⁹ package and the Codonml routine from PAML.⁵⁰ In order to analyze the yeast data, 1,000 independent instances for two null models, *Tree Shuffle* and *Network Shuffle* (as detailed in Figure 8.2), were generated. These randomly reassign phylogenies to nodes in the network, and rewire the network while keeping the degree of each node fixed, respectively.

Phylogenetic trees for each protein were inferred by first aligning each protein sequence with its available orthologues in the other yeast species. These multi-sequence alignments were then used to infer the topology of the evolutionary relationship. Three different algorithms were used to infer trees: we used the PARS and PROML programmes of the Phylip 3.6⁴⁹ package, and the Codonml routine from PAML.⁵⁰ In order to compare the results for the different inferential procedures

we have to take into account that PARS generates bifurcating trees, while the two maximum likelihood approaches (henceforth denoted by PROML and PAML) infer multifurcating tree structures.

Crucially, the topologies of the gene trees can differ from the presumed species tree. To examine the similarity of phylogenetic trees, the number of possible tree shapes for each method of tree construction is of critical importance and a potentially confounding factor in the analysis. In the following study, rooted trees are considered, created using *bifurcating* and *multifurcating* methodologies. *Bifurcating* trees are defined as those where every interior node is of degree 3, whilst every tip is of degree 1 (only connecting to one other ancestral node). *Multifurcating* trees, on the other hand, can have interior nodes with a higher degree, increasing the possible number of topologies available for a fixed number of sequences (the set of all *multifurcating* trees also contains all *bifurcating* trees).

We restrict our analysis to those proteins for which trees can be inferred unambiguously. This differs slightly between the different methods and therefore the number of comparisons differs across phylogeny inference procedures. For each method the number of homologues found, on average, for each phylogenetic tree is above five. Given two trees, their shapes are defined as matching if the trees, on the restricted subset of shared species, are identical. Clearly, a minimum tree size is needed for a match (if they share less than three species the trees will always match), and we therefore only consider cases where at least three shared species appear in the two phylogenies. A match means that in the set of species which are used in the comparison, inferred phylogenies reveal no mismatch.

When looking to compare the similarity of phylogenetic trees, strict identity is a conservative measure, especially when the proteins share a large number of homologues across the yeast study species. To augment this simple and coarse measure we assess how different the trees of interacting proteins are. This method allows the comparison of non-matching pairs of phylogenetic trees. Our approach for measuring similarity between trees is based on a nearest-neighbour interchange method. A neighbour is defined as any tree that can be reached by moving a particular lineage either inside or outside of a neighbouring internal node. In the case of a bracketed

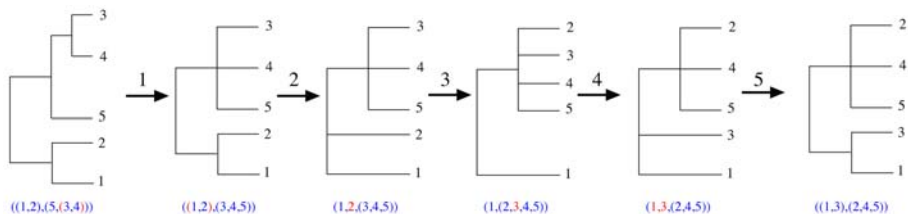


Fig. 8.4. An example showing how the scoring function works between different phylogenetic topologies.

tree representation (see *e.g.* Fig. 8.4) this means that a species is moved across one of the two nearest brackets specifying the topology. The score, $s_{a,b}$, is the minimum number of such moves necessary until the two trees, of proteins a and b , match. The scheme searches the space of neighbours and reports the minimal number of branch swaps between the two trees, using the space of multifurcating topologies as the search space between trees.

In order to be able to compare the scores over different numbers of homologous proteins, a further scoring function is used across each dataset. This is necessary as the space of possible topologies is different depending on the number of shared homologues, so the scores are not directly comparable across different numbers of shared homologues. This score, $E_{a,b}$ for proteins a and b , gives a score in $[0, 1]$ – the higher the value the closer the match between the topologies in question. The score takes into account the number of possible moves between the two topologies, which is dependent on M_n – the number of possible topologies for trees on n species. Accordingly, we define $E_{a,b} = 1 - \frac{s_{a,b}}{M_n}$, where $s_{a,b}$ is the score between the two trees sharing n species and M_n is the maximum possible score between two trees on n species.

8.4.2. Coevolution in phylogenies: BC confidence intervals

Results obtained for basic topology matches across interacting pairs are summarized independently for the two statistical null models in Table 8.1 and Table 8.2. We have employed three different phylogenetic algorithms and analyzed three PPI datasets. We find identical trends for the different phylogenetic algorithms. However, the proportion of detected matches recorded for the real PIN data varies considerably across the different methods. For example, in the case of the CORE network data, phylogenies inferred using PAML match in approximately 17% comparisons, phylogenies inferred using PROML match in 42% and phylogenies inferred using PARS match in 57%. These differences can be explained by the difference in complexity of both the possible number of *bifurcating* and *multifurcating* topologies, as well as

Table 8.1. The percentage of matching topologies and average score per comparison for phylogenies inferred using phylogeny methods on different protein interaction datasets are shown together with the results of the *Network Shuffle* null model.

Method	Data	Real (%)	> (%)	Net Shuffle Match $\hat{\mu}$ [p0.05, p0.95]	Match Score	> (%)	Net Shuffle Score $\hat{\mu}$ [p0.05, p0.95]
PAML	CORE	16.7	88.0	17.3 [16.5, 18.1]	0.703	29.2	0.702 [0.697, 0.707]
	DIP	16.3	100.0	17.5 [17.0, 18.0]	0.703	99.6	0.707 [0.704, 0.710]
	LC	16.0	100.0	16.9 [16.5, 17.3]	0.701	75.7	0.702 [0.700, 0.705]
PROML	CORE	41.5	98.3	42.7 [41.8, 43.6]	0.835	18.4	0.836 [0.833, 0.840]
	DIP	39.1	100.0	40.1 [39.6, 40.6]	0.829	70.2	0.830 [0.828, 0.831]
	LC	39.0	99.8	39.8 [39.4, 40.3]	0.829	78.3	0.828 [0.827, 0.830]
PARS	CORE	56.9	86.6	57.8 [56.5, 59.0]	0.888	90.8	0.891 [0.887, 0.895]
	DIP	55.4	81.9	55.8 [55.1, 56.5]	0.885	73.2	0.886 [0.884, 0.888]
	LC	54.7	96.7	55.4 [54.8, 56.0]	0.884	75.9	0.885 [0.883, 0.887]

Table 8.2. The percentage of matching topologies and average score per comparison for phylogenies inferred using phylogeny methods on different protein interaction datasets are shown together with the results of the *Node Shuffle* null model.

Method	Data	Real	>	<i>Node Shuffle</i> Match		Real	>	<i>Node Shuffle</i> Score	
		(%)	(%)	$\hat{\mu}$	$[p_{0.05}, p_{0.95}]$	Score	(%)	$\hat{\mu}$	$[p_{0.05}, p_{0.95}]$
PAML	CORE	16.7	5.6	15.2	[13.7, 16.8]	0.703	17.4	0.697	[0.686, 0.708]
	DIP	16.3	2.9	15.2	[14.2, 16.2]	0.703	11.6	0.697	[0.688, 0.707]
	LC	16.0	11.8	15.2	[14.1, 16.3]	0.701	18.3	0.696	[0.688, 0.705]
PROML	CORE	41.5	7.5	39.6	[37.3, 41.9]	0.835	1.9	0.823	[0.814, 0.833]
	DIP	39.1	57.1	39.2	[36.5, 41.1]	0.829	3.4	0.822	[0.807, 0.831]
	LC	39.0	67.2	39.6	[37.7, 41.5]	0.829	11.1	0.823	[0.815, 0.831]
PARS	CORE	56.9	0.7	53.2	[50.7, 55.9]	0.888	0.5	0.874	[0.864, 0.884]
	DIP	55.4	4.2	53.3	[51.3, 55.4]	0.885	0.8	0.874	[0.865, 0.882]
	LC	54.7	14.7	53.4	[51.3, 55.5]	0.884	2.2	0.875	[0.866, 0.882]

differences in the construction methods.

Table 8.2 clearly indicates that there are more topology matches between interacting proteins, on average, in the true network data than in the *Node Shuffle* null model replicates, except in the case of PROML where the true average is close to the *Node Shuffle* results. Moreover, as the network considered changes (from CORE to LC), the experimental data shows a lower proportion of matching topologies, while the *Node Shuffle* results stay constant across the construction approaches.

Under the *Network Shuffle* null model topologies match more frequently by chance than in the true data, as shown in Table 8.2. This null model fixes the degree associated with each gene-tree, resulting in more topology matches from the random networks. This reflects the importance of the gene trees of the hub proteins (highly connected proteins) in network analyses. Thus the hubs appear to be more similar to a random protein than to their reported interaction partners. Figure 8.2 (b) shows the relative proportions of matching gene trees for different numbers of shared homologues for the DIP data (as this determines the number of possible topologies, and accordingly the probability of a match of random phylogenies).

Splitting the data by the number of homologues compared shows differences between the tree construction methods. In the PAML case, shown in panel (c) of Fig. 8.3, for a fixed number of homologues compared, the scores are higher than those obtained from the second maximum likelihood method, PROML. However both methods show the same trend across the different numbers of species included in the comparison. Indeed, the main discrepancy gleaned from the mismatch scores is caused by the maximum parsimony method, PARS, which generates *bifurcating* phylogenies while the scoring function is based on *multifurcating* trees. Finally, a phylogeny with fewer species will naturally tend to produce more matches and lower mismatch scores than one with more species. The average match results are confirmed with the further analysis using the scoring function detailed in *Methods*.

The *Tree Shuffle* null model suggests that topologies in the true data are more similar, whereas the *Network Shuffle* null model shows that random allocations into interacting pairs provide a higher average score across all the comparisons. The CORE data – seen in Table 8.1 – has the most significant evidence of more

similarity in the real data (for the maximum likelihood inference methods), although the results are not statistically significant (even for a 10% one-sided hypothesis test). Every possible protein pair was compared to see how similar the tree structures were over the whole space of possible interactions. For every possible protein pair, the proportion of matches were: 40% (PROML), 56% (PARS), 15% (PAML). These results are lower than in the true network data.

It seems that in *S. cerevisiae* we cannot use a reported match of the topologies of two proteins to infer protein interactions with high reliability. Indeed, in our already quite extensive dataset there appears to be a slightly negative correlation, as random networks (i.e. keeping the phylogeny associated with a node of certain degree and randomly rewiring the edges) appear to have more protein pairs with matching topologies. These results concerning the topology of interacting proteins do not, however, necessarily contradict previous work on coevolution of interacting proteins.^{3,38,44,46} Measures of the evolutionary rate or functional similarity are not accounted for in this analysis and could easily correlate with interactions; in yeast (and also in *C. elegans*), however, there is evidence that such a correlation among the evolutionary rates on interacting proteins is at best weak.⁴

8.4.3. *Coevolution measured by rates: conditioning on additional data*

Figure 8.5 shows the correlations, measured using Kendall's τ rank correlation statistic, between the evolutionary rates of interacting proteins (observed values are indicated by vertical red lines) in the *S. cerevisiae* PIN. Histograms resulting from the BC null model (black) and null models using `GOcardShuffle` with one (red), two (green) and three (blue) gene ontology categories are also shown in the same figure. Under the BC null model the evolutionary rates of interacting proteins appear to be significantly correlated. The histograms of the conditional Null models move further towards the observed values of τ as more GO information is being included into the null model. Using the full annotation results in a histogram (or ensemble of conditional networks) which covers the observed correlation among evolutionary rates of interacting proteins.

We also observe that different GO annotations appear to correlate to different extents with the evolutionary rate. Functional annotations appear to have a greater effect in explaining variation in evolutionary rates than process annotations. The cellular component annotations, finally, explain very little of the variation in evolutionary rates. This agrees with earlier results.^{4,37}

8.5. Network Analysis and Confounding Factors

We have shown above that it is possible to tune null models for network organization that are based on conventional BC graphs such that the networks from the conditional ensemble also reflect other properties of the true network. These prop-

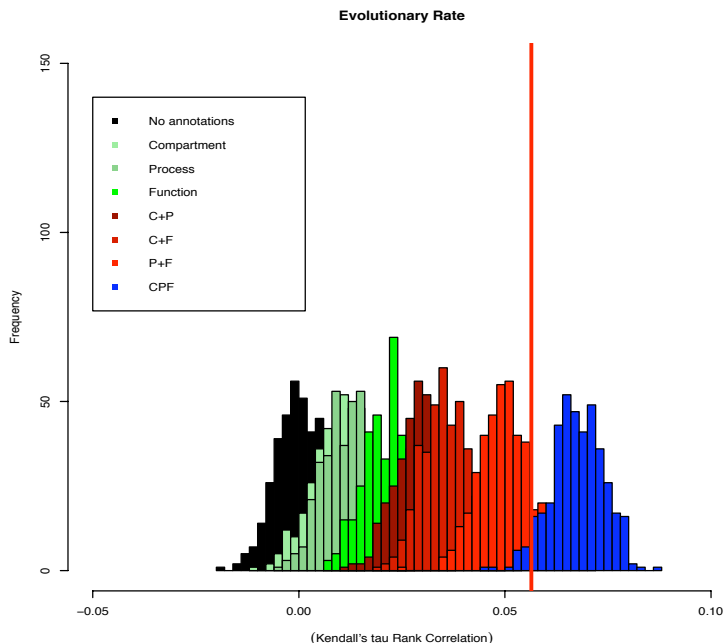


Fig. 8.5. Confidence intervals for the correlation of evolutionary rates among pairs of interacting proteins. The real data is indicated by a red vertical line. Incorporating GO annotations, individually, in pairs, or all three categories together results in progressive right-shifts of the distribution under the conditional Null models. Function, Process and Compartment are indicated by F, P and C, respectively.

erties may include other network statistics on top of the degree sequence, such as the clustering coefficient or the degree-degree distribution. Alternatively, we may want to include other co-variate data which may reflect higher levels of organization in the network. The gene ontology information, which can be captured by `GOcardShuffle` as shown above.

Two points are worth noting and reiterating: if we always reject a null hypothesis then this should suggest to us that the null hypothesis is wrong or inadequate. We have seen this repeatedly in network analyses, where properties of pairs of interacting proteins, for instance, were sufficiently more similar than was expected to occur by chance. Chance here refers implicitly to the properties of an ensemble of BC networks. The persistence with which these observations appear in the literature is precisely the reason why we should go beyond simple BC graphs as Null models of network organization (although, as the example of phylogenies discussed above shows, for sufficiently weak or spurious signals, even the BC ensemble may include observed correlations among the properties of interacting proteins).

The second and intimately related point relates to the confounding nature of net-

works in any statistical analysis. In statistics we refer to situations where inclusion of a confounding (or hidden or lurking) variable alters or reverses the correlation between different variables as an example of Simpson's paradox: this occurs when the correlation between two random vectors A and B , $c(A, B)$, is different in nature compared to the correlation conditional on some other random vector, C , $c(A, B|C)$. If there are any higher levels of organization in the network than the mere connectivity patterns among nodes, then these will act as global confounding factors. In a cellular context such hierarchical organization will be omnipresent: proteins in the mitochondria will interact predominantly with other mitochondrial proteins, ribosomal proteins with other ribosomal proteins etc.. If we ignore this coarse-grained structure of biological networks, then we may fall foul of Simpson's paradox and detect spurious associations.

These factors, unfortunately, conspire against straightforward evolutionary analysis: the statistical inference of parameters will be far from trivial, and the mathematical models used to model network evolution are far from realistic. In a non-parametric manner it is, however, possible to incorporate additional biological or genomic data into the statistical analysis of biological systems as we have argued. This in turn can help us in identifying the principal factors underlying network organization, and hopefully, network evolution.

References

1. E. Alm and A.P. Arkin, Biological networks. *Curr. Opin. Struct. Biol.* **13**, 193–202, (2003).
2. E. de Silva and M.P.H. Stumpf, Complex networks and simple models in biology. *J.Roy.Soc. Interface.* **2**, 419–340, (2005).
3. C.S. Goh and F.E. Cohen, Co-evolutionary analysis reveals insights into protein-protein interactions. *J. Mol. Biol.* **324**, 177–192, (2002).
4. I. Agraftioti, J. Swire, J. Abbott, D. Huntley, S. Butcher and M.P.H. Stumpf, Comparative analysis of the *saccharomyces cerevisiae* and *caenorhabditis elegans* protein interaction networks. *BMC Evolutionary Biology*. bf 5, 23, (2005).
5. L. Hakes, S.C. Lovell, S.G. Oliver and D.L. Robertson, Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl. Acad. Sci. USA.* **104**, 7999–8004, (2007).
6. J. Felsenstein, *Inferring Phylogenies*. Sinauer Associates, (2003).
7. I. Xenarios, D. Rice, L. Salwinski, M. Baron, E. Marcotte, and D. Eisenberg, Dip: the database of interacting proteins. *Nucl. Acid. Res.*, **28**, 289–291, (2000).
8. H. Hermjakob, L. Montecchi-Palazzi, G. Bader, R. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, L. Xenarios, D. Eisenberg, B. Steipe, C. Hogue and R. Apweiler, The hupopsi's molecular interaction format - a community standard for the representation of protein interaction data. *Nature Biotech.* **22**, 177–183, (2004).
9. M. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, J. Schultz, J. Rick, A.

- Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, M. Hudak, D. Dickson, T. Rudi, V. Ganu, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E.V.R. Querfurth, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. **415**, 141–147, (2002).
10. E. de Silva, T. Thorne, P. Ingram, I. Agraftioti, J. Swire, C. Wiuf and M.P.H. Stumpf, The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biology*. **4**, 39, (2006).
 11. M.P.H. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. An, M. Lappe and C. Wiuf, From the cover: Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA*. **105**, 6959–6964, (2008).
 12. S. Silvey, *Statistical Inference*. Chapman & Hall, (1975).
 13. M. Middendorff, Z. Etay and C. Wiggins, Inferring network mechanisms: The drosophila melanogaster protein interaction network. *Proc. Natl. Acad. Sci. USA*. **102** 3192–3197, (2005).
 14. O. Ratmann, O. Jorgensen, T. Hinkley, M.P.H. Stumpf, S. Richardson and C. Wiuf, Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *h. pylori* and *p. falciparum*. *PLoS Comput. Biol.* **3**, 2266–2278, (2007).
 15. M.P.H. Stumpf, W.P. Kelly, T. Thorne and C. Wiuf, Evolution at the system level: the natural history of protein interaction networks. *Trends Ecol.Evol.* **22**, 366–373, (2007).
 16. A. Krzywicki, Defining statistical ensembles of random graphs. *arXiv cond-mat*. 0110574, (2001).
 17. M. Newman, The structure and function of networks. *Comp. Phys. Comm.* **147**, 40–45, (2002).
 18. S. Dorogovtsev and J. Mendes, *Evolution of Networks*. Oxford University Press, (2003).
 19. B. Bollobás and O. Riordan, Mathematical results on scale-free graphs. In S Bornholdt and H Schuster, editors, *Handbook of Graphs and Networks*, 1–34. Wiley-VCH, (2003).
 20. P. Erdős and A. Rényi, On random graphs. *Publicaciones Mathematicae Debrecen*. **5**, 290–297, (1959).
 21. P. Erdős and A. Rényi, On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5**, 17–61, (1960).
 22. E. Gilbert, Random graphs. *Ann. of Math.Stats.* **30**, 1141–1144, (1959).
 23. E. Bender and E. Canfield, The asymptotic number of labeled graphs with given degree sequence. *J. Comb. Theory A*. **24**, 296–307, (1978).
 24. M. Molloy and B. Reed, A critical point for random graphs with a given degree distribution. *Rand. Struct. Algorithms*. **6**, 161–179, (1995).
 25. M. Molloy and B. Reed, The size of the giant component of a random graph with a given degree sequence. *Comb. Probab. Comput.* **7**, 295–305, (1998).
 26. N. Newman, S. Strogatz and D. Watts, Random graphs with arbitrary degree distributions and their applications. *Phys.Rev. E*. **64**, 026118, (2001).
 27. M. Newman, Random graphs as models of networks. In S Bornholdt and H Schuster, editors, *Handbook of Graphs and Networks*. Wiley-VCH, (2003).
 28. N. van Kampen, *Stochastic Processes in Physics and Chemistry*. North-Holland, (1992).
 29. C. Wiuf, M. Brameier, O. Hagberg and M.P.H. Stumpf, A likelihood approach to the analysis of network data. *Proc. Natl. Acad. Sci. USA*, **103**, 7566–7570, (2006).
 30. T. Thorne and M.P.H. Stumpf, Generating confidence intervals on biological networks.

- BMC Bioinformatics*. **8**, 467, (2007).
31. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092, (1953).
 32. B.D. Ripley, *Stochastic Simulation*. Wiley, (1987).
 33. C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2nd edition, (2004).
 34. M. Newman and G. Barkema, *Monte Carlo Methods in Statistical Physics*. Clarendon Press, (1999).
 35. H.B. Fraser, A.E. Hirsh, L.M. Steinmetz, C. Scharfe and M.M. Feldman, Evolutionary rate in the protein interaction network. *Science*. **296**, 750–752, (2002).
 36. I.K. Jordan, Y.I. Wolf and E.V. Koonin, No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. *BMC Evol. Biol.* **3**, 1, (2003).
 37. D. Drummond, A. Raval and C. Wilke, A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23**, 327–337, (2006).
 38. C.S. Goh, A.A. Bogan, M. Joachimiak, D. Walther and F.E. Cohen, Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293, (2000).
 39. J. Gertz, G. Elfond, A. Shustrova, M. Weisinger, M. Pellegrini, S. Cokus and B. Rothschild, Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*. **19**, 2039–2045, (2003).
 40. A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol.Biol.Evol.* **18**, 1283–1292, (2001).
 41. J. Yu and F. Fotouhi, Computational approaches for predicting protein-protein interactions: A survey. *J. Med. Sys.* **30**, 39–44, (2006).
 42. M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg and T. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U S A*. **96**, 4285–8, (1999).
 43. F. Pazos and A. Valencia, Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*. **14**, 609–614, (2001).
 44. F. Pazos, J. Ranea, D. Juan and M.J.E. Sternberg, Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.* **352**, 1002–15, (2005).
 45. T. Sato, Y. Yamanishi, M. Kanehisa and H. Toh, The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*. **21**, 3482–3489, (2005).
 46. A. Ramani and E. Marcotte, Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–84, (2003).
 47. K. Wolfe, Comparative genomics and genome evolution in yeast. *Phil. Trans. Roy. Soc. Lond. B. Biol.Sci.* **361**, 403–412, (2006).
 48. D. Fitzpatrick, M. Logue, J. Stajich and G. Butler, A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.* **6**, 99, (2006).
 49. J. Felsenstein, Phylip - phylogeny inference package (version 3.2). *Cladistics*. **5**, 164–166, (1989).
 50. Z. Yang, Paml: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in Biosciences*. **13**, 555–556, (1997).

This page intentionally left blank

Index

- 16SrRNA, 135
- NP*-complete, 13, 50

- adjacency matrix, 9, 69, 81, 151
- algorithm, 11, 31, 45, 50, 52, 75, 101, 131, 146, 154, 157
- annotation, 151, 153, 161
- approximate Bayesian computation, 31
- Arabidopsis thaliana*, 127
- architecture, 50, 55
- ATP, 115–117
- average path, 9
 - length, 8, 90, 91

- basic reproduction number, 98
- Bayesian inference, 27
- Bender-Canfield (BC) network, 149
- betweenness, 103
- bifurcating, 158
- binding site, 65, 76
- biological process, 27, 45, 145, 153
- Black Death, 89
- BLAST, 130, 131
- Boltzmann, Ludwig, 148
- bond percolation, 104
- Brownian motion, 131
- building block, 46
- burn-in period, 155

- Caenorhabditis elegans*, 18, 68, 128
- cancer, 68
- canonical ensemble, 148
- cell cycle, 54
- chemokine, 156
- ChIP-on-chip, 68
- chromatin immunoprecipitation (ChIP), 68
- classification, 56
- cluster, 49, 56, 60, 66, 70, 73, 78, 95, 105

- clustering coefficient, 7, 92, 93, 150, 162
- coevolution, 129, 134, 137, 147, 155
- community, 91, 153
- compartmental model, 85, 88
- complexity, 2, 11, 14, 21, 28, 45, 65, 94, 120, 127, 151, 159
- confidence interval, 151
- connected component, 4, 9, 30, 104
- connectivity, 22, 56, 66, 70, 119, 120, 128
 - pattern, 163
- conservation, 19, 61, 76, 78, 130, 155
- contact network, 2, 91, 102, 103
- control coefficient, 117
- coregulation, 55
- correlated mutation, 130, 133
- correlation, 55, 66, 73, 76, 95, 108, 134, 136
- cortical network, 55

- Drosophila melanogaster*, 21, 59, 60, 68, 128
- database, 146
- degree, 4, 23, 26, 51, 86, 153
 - distribution, 6, 21, 22, 24, 34, 51, 89, 90, 99, 107, 108, 150, 153
 - sequence, 21, 25, 29, 148, 149
- density dependent, 87
- design pattern, 46
- diameter, 5, 9, 30, 90
- DIP data, 160
- disease, v, 86, 90, 103
 - transmission, 99
- distance, 5, 12, 30, 31, 56, 90, 130, 132, 134, 156
- divergence, 12, 23, 34, 76, 130
- DNA sequence, 47, 53
- domain, 34, 65, 130, 131, 134, 146
- dominance, 118

- E-value, 132
- ecological and epidemiological interaction, 2
- electrophoresis, 68
- Elementary Modes Analysis, 113, 114
- emerging network, 121
- Ensembl database, 78
- ensemble, 50, 59, 69, 145, 148
- enzyme, 46, 76, 113, 118, 120, 132
- epidemic, 87, 99
- epistasis, 115, 117
- equilibrium, 88
- Erdős–Rényi (ER), 56, 90
 - graph, 21, 149
 - model, 70
- Escherichia coli*, 12, 48, 53, 74, 114
- eukaryote, 17, 20, 23
- evolution, v, 2, 17, 18, 20, 51, 66, 79, 113, 116, 117, 121, 128, 137
- evolutionary, 146
 - conservation, 17, 76
 - dynamics, 2, 27, 66
 - game, 117
 - process, 2, 12, 23, 113, 127, 137, 150
- experimental protocol, 20
- Exponential Random Graph Model (ERGM), 21
- expression level, 46, 66, 128, 151, 156
- false negative, 68, 147
- false positive, 68, 147
- fitness, 117
- flux, 114
- Flux Balance Analysis, 113
- food web, 2, 99
- foot-and-mouth disease, 89
- fragmentation, 30
- frequency concept, 49
- frequency dependent, 89
- functional unit, 61, 66
- fuzziness, 74
- gene
 - duplication, 17, 18, 33, 79, 121
 - expression, 53, 68, 118, 134, 138
 - fusion, 133, 138
 - neighboring, 130, 138
 - ontology (GO), 153, 161
 - regulation network, 1, 57, 113, 122
- genome, 66, 131
 - genome-wide scale, 65
 - giant connected component (GCC), 6, 149
 - Gilbert, Edgar N., 149
 - GOcardShuffle, 153
 - graph alignment, 73
 - Gravisto, 52
 - Helicobacter pylori*, 17, 128
 - Hamming distance, 74, 132
 - heterogeneity, 87, 96
 - high-confidence, 20
 - high-throughput, 127
 - technology, 45
 - HIV, 108
 - homeostasis, 59
 - Homo sapiens*, 65, 78
 - homologue, 130, 156
 - horizontal gene transfer (HGT), 135
 - hot spot, 129
 - hub, 119, 120, 129, 160
 - in-degree, 4
 - incompleteness, 20
 - infection, 85
 - IntAct, 21
 - interactome, 20, 21, 127, 138, 147
 - Internet, 99
 - isomorphic, 12, 28, 46
 - Keeling clustered network, 94
 - Kendall's τ rank correlation, 161
 - Kermack–McKendrick model, 85
 - kinetics, 87, 116, 122
 - knockout, 121
 - mutation, 114
 - lateral gene transfer, 18
 - lattice, 6, 20, 56, 105
 - lethal, 118
 - likelihood, 27, 28, 154, 160
 - likelihood-free inference, 18, 31
 - log-likelihood, 71, 76
 - loop, 3, 4, 73, 93, 146
 - Lynch, Michael, 2
 - Mus musculus*, 78
 - macroscopic, 148
 - malaria, 85
 - Markov chain, 23, 151, 155

- Markov Chain Monte Carlo (MCMC), 31, 154
- mass spectrometry, 128
- mass-action, 87
- master equation, 24, 151
- match, 47
- Matthew effect, 101
- MAVisto, 52
- maximum likelihood, 71, 158
- Mcm1, 57
- mean-field, 87
- measles, 87
- mesoscopic system, 11
- metabolic
 - network, 2
 - pathway, 115
- Metabolic Control Analysis (MCA), 114
- metabolite, 46, 76, 113, 117, 120
- metabolome, 18
- Metropolis sampling, 154
- Mfinder, 52
- microarray, 68, 78
- microscopic state, 147, 148
- Molloy-Reed criterion, 149
- moment closure, 91, 94, 96
- motif, 45
 - bi-fan, 45, 46, 56
 - feed-forward loop motif, 46, 47, 56
 - fingerprint, 54
 - multi-input, 46
 - single-input, 46
- mRNA, 73
- multicellular, 19, 20, 128
- mutation, 19, 117, 118, 121, 136
- neighbour, 4, 7, 23, 91, 98, 158
- neighbourhood, 4, 87, 95, 152
- neo-functionalisation, 19
- NetShuffle, 152, 153
- network
 - evolution, 2, 17, 27, 35, 56, 80, 113, 147, 163
 - growth, 18, 23, 27, 59, 151
 - theme, 57
- neural
 - net, 137
 - synapse, 57
- neutral evolutionary theory, 2
- node centrality, 103
- NodeShuffle, 152
- noise, 20, 29, 58
- non-functionalisation, 19, 79
- null
 - hypothesis, 50, 60, 80, 147
 - model, 21, 51, 58, 69, 71, 77, 145, 147, 157, 159
- open reading frame, 21, 29
- operon, 130, 138
- optimal design, 115
- order, 30, 31
- organization, 147, 150, 152, 161
- orthologue, 65, 66, 77, 130, 133, 157
- out-degree, 4, 6, 87
- P-value, 51, 53
- Plasmodium falciparum*, 17, 19
- pairwise mismatch, 73
- Pajek, 52
- PAML, 157, 159
- path, 4, 5, 150
- pattern, 13, 14, 18, 23, 50, 57, 59, 69–71, 73, 76, 81, 109, 117, 130
- Pearson correlation, 151
- percolation threshold, 96, 105
- permutation, 152
- phosphorylation, 113
- Phylip, 157
- phylogenetic, 18, 61, 130–132, 135, 146, 156–158
- plasticity, 19, 59, 80
- Poisson
 - distribution, 69, 99
 - random network, 90, 91, 97
- posterior, 33, 34, 80, 81
 - density, 27
- power-law, 22, 23, 120
- preferential attachment, 101, 102, 120
- prior, 27, 80, 81
- prokaryote, 17, 23, 130
- promoter, 47, 130
- protein interaction network (PIN), 2, 12, 14, 17, 59, 68, 70, 72, 128, 138, 146
- protein-DNA interaction, 65
- protein-protein interaction network, 156
- proteome, 127, 128, 135
- pyridoxine, 73
- random
 - graph, 20, 90, 91, 99, 145, 153

- network, 69, 93, 100, 108, 148, 150, 160
- Randomly Grown Graph (RGG), 22
- receptor, 134, 156
- recombination, 118
- regulon, 53
- reticulation, 146
- rewiring, 21, 51, 91, 151, 153, 161
- ribosomal protein, 163
- robustness, 115, 117, 120, 121
- Saccharomyces cerevisiae*, 12, 19, 32, 34, 46, 53, 54, 57, 65, 147, 153, 156, 161
- sampling
 - bias, 20, 29, 33
 - fraction, 21, 29
- selection, 2, 50, 73, 79, 80, 115, 120, 124, 138
- sequence, 113
 - alignment, 133, 134
 - similarity, 76
- sexually transmitted infection (STI), 97
- shortcut, 91, 105
- signal transduction, 57, 58, 60, 113, 122
- signalling cascade, 128
- similarity, 76
- Simpson's paradox, 163
- single gene duplication, 18
- SIR*, 85, 94, 102
- Sir Ronald Ross, 85
- site percolation, 104
- size, 30
- small-world network, 88, 91, 105
- spanning tree, 5
- stoichiometry, 114
- Strogatz, Steven H., 91
- structural stability score, 58
- sub-functionalisation, 19, 79
- summary statistic, 30, 32
- supernode, 105
- supervised method, 137
- susceptibility, 105
- SVM, 137
- Swi4, 57
- Szathmáry, Eörs, 118
- Treponema pallidum*, 17, 19, 34
- thinning-out interval, 155
- topology, 6, 18, 28, 51, 56, 65, 73, 128, 156, 159, 160
- transcription factor, 17, 57, 68, 76
 - binding site, 12
- transcriptional network, 2
- transitivity, 8
- transmission, 85, 91, 97, 98, 103, 107
- Tree Shuffle, 157
 - null model, 152
- tree-like, 74
- triad, 52
- Tryptophan operon, 130
- Uetz, Peter, 73
- undirected and directed graphs, 49
- unicellular, 19, 20, 128
- variance-to-mean, 98
- Voronoi tessellation, 95
- Watts, Duncan J., 91
- whole genome duplication, 18, 151
- within-reach distribution, 30
- World Wide Web, 99, 104
- yeast, 46, 53, 54, 57, 65, 68, 73, 157
- yeast two-hybrid (Y2H), 68
- yield, 115
- Z-score, 51, 53, 60