



# Networks in Cell Biology

EDITED BY

Mark Buchanan, Guido Caldarelli,  
Paolo De Los Rios, Francesco Rao  
and Michele Vendruscolo

CAMBRIDGE

CAMBRIDGE

[www.cambridge.org/9780521882736](http://www.cambridge.org/9780521882736)

This page intentionally left blank

## NETWORKS IN CELL BIOLOGY

The science of complex biological networks is transforming research in areas ranging from evolutionary biology to medicine. This is the first book on the subject, providing a comprehensive introduction to complex network science and its biological applications.

With contributions from key leaders in both network theory and modern cell biology, this book discusses the network science that is increasingly foundational for systems biology and for the quantitative understanding of living systems. It surveys studies in the quantitative structure and dynamics of genetic regulatory networks, molecular networks underlying cellular metabolism, and other fundamental biological processes. The book balances empirical studies and theory to give a unified overview of this interdisciplinary science. It is a key introductory text for graduate students and researchers in physics, biology and biochemistry, and presents ideas and techniques from fields outside the reader's own area of specialization.

MARK BUCHANAN is a physicist and independent author. He writes a monthly column for the journal *Nature Physics*.

GUIDO CALDARELLI is Associate Professor in the Centre for Statistical Mechanics, CNR-INFM, Rome, and collaborates with the Research Centre 'E. Fermi', the Laboratory LINKALAB and the London Institute for Mathematical Sciences.

PAOLO DE LOS RIOS is a Professor in the Laboratory of Statistical Biophysics, Institute of Theoretical Physics, Ecole Polytechnique Federale de Lausanne (EPFL).

FRANCESCO RAO researches in the Laboratoire de Chimie Biophysique at the University of Strasbourg and at the Research Centre 'E. Fermi', Rome.

MICHELE VENDRUSCOLO is a Reader in Theoretical Chemical Biology in the Department of Chemistry at the University of Cambridge.



# Networks in Cell Biology

*Edited by*  
M. BUCHANAN

G. CALDARELLI  
*CNR-INFM*

P. DE LOS RIOS  
*Ecole Polytechnique Federale de Lausanne*

F. RAO  
*University of Strasbourg*

M. VENDRUSCOLO  
*University of Cambridge*



**CAMBRIDGE**  
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS  
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,  
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press  
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

[www.cambridge.org](http://www.cambridge.org)

Information on this title: [www.cambridge.org/9780521882736](http://www.cambridge.org/9780521882736)

© Cambridge University Press 2010

This publication is in copyright. Subject to statutory exception and to the provision of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published in print format 2010

ISBN-13 978-0-511-73118-1 eBook (Adobe Reader)

ISBN-13 978-0-521-88273-6 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of urls for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

# Contents

<i>List of contributors</i>	page viii
Introduction	1
<i>M. Buchanan, G. Caldarelli, P. De Los Rios, F. Rao and M. Vendruscolo</i>	
1 Network views of the cell	4
<i>Paolo De Los Rios and Michele Vendruscolo</i>	
1.1 The network hypothesis	4
1.2 The central dogma and gene regulatory networks	5
1.3 Protein–protein interaction networks	7
1.4 Metabolic networks	9
1.5 Signaling networks	11
1.6 Networked networks and cell functionality	12
1.7 Concluding remarks	13
2 Transcriptional regulatory networks	14
<i>Sarath Chandra Janga and M. Madan Babu</i>	
2.1 Introduction	14
2.2 Transcriptional regulation in prokaryotes and eukaryotes	19
2.3 Structure of transcriptional regulatory networks	20
2.4 Evolution of transcriptional regulatory networks	25
2.5 Dynamics of transcriptional regulatory networks	29
2.6 Conclusions	34
3 Transcription factors and gene regulatory networks	36
<i>Matteo Brilli, Elisa Calistri and Pietro Lió</i>	
3.1 Introduction	36
3.2 Promoters' complexity/eukaryotic gene promoters	37
3.3 Transcription factors	41
3.4 Bioinformatics of regulatory networks	46

4	Experimental methods for protein interaction identification	53
	<i>Peter Uetz, Björn Titz, Seesandra V. Rajagopala and Gerard Cagney</i>	
4.1	Introduction	53
4.2	Protein complementation techniques	57
4.3	Affinity purification methods	62
4.4	Protein complex purification and mass spectrometry	65
4.5	Protein and peptide chips	73
4.6	Other methods for interaction detection and functional analysis	74
4.7	Quality of large-scale interaction data	75
4.8	Comparison of methods	77
4.9	Conclusions	80
5	Modeling protein interaction networks	83
	<i>Francesco Rao</i>	
5.1	Introduction	83
5.2	Scaling laws and network topology	85
5.3	Predicting protein interactions	87
5.4	Towards models at an atomic level of resolution	88
5.5	Concluding remarks	90
6	Dynamics and evolution of metabolic networks	93
	<i>Daniel Segrè</i>	
6.1	Introduction	93
6.2	Cellular metabolism and its regulation	93
6.3	Metabolism across disciplines	96
6.4	Dynamics of a metabolic system	98
6.5	Stoichiometric analysis	100
6.6	Constraint-based modeling: feasible states and optimality	103
6.7	Predicting genetic perturbations	108
6.8	Double perturbations and epistatic interactions	109
6.9	The ancient history of metabolism: from cell-scale to biosphere-scale	112
6.10	Conclusions	116
7	Hierarchical modularity in biological networks: the case of metabolic networks	117
	<i>Erzsébet Ravasz, Regan</i>	
7.1	Introduction	117
7.2	Modularity and hubs in biological networks	119
7.3	Scaling of the clustering coefficient: a signature of hierarchy	121
7.4	Method for finding network modules	123
7.5	A case study: the <i>E. coli</i> metabolic network	126



7.6	Hierarchy, fractality and the small world of networks	131
7.7	Conclusions	134
8	Signalling networks	135
	<i>Gian Paolo Rossini</i>	
8.1	Introduction	135
8.2	Chemical signalling: many pathways following a few general themes	136
8.3	Cross-talks among signal transduction pathways	159
8.4	Signalling networks, system organization and modelling	166
8.5	Conclusions and outlook	169
Appendix A	Complex networks: from local to global properties	170
	<i>D. Garlaschelli and G. Caldarelli</i>	
Appendix B	Modelling the local structure of networks	188
	<i>D. Garlaschelli and G. Caldarelli</i>	
Appendix C	Higher-order topological properties	210
	<i>S. Ahnert, T. Fink and G. Caldarelli</i>	
Appendix D	Elementary mathematical concepts	219
	<i>A. Gabrielli and G. Caldarelli</i>	
	<i>References</i>	235
	<i>Index</i>	269

## Contributors

S. Ahnert  
TCM Group Cavendish Laboratory  
University of Cambridge  
Cambridge  
CB3 0HB  
UK

M. Madan Babu  
MRC-Laboratory of Molecular Biology  
University of Cambridge  
Hills Road  
Cambridge  
CB2 0QH  
UK

Matteo Brilli  
Biométrie et Biologie Evolutive  
UMR CNRS 5558  
Université Lyon 1  
43, Bvd du 11 Novembre  
69622 Villeurbanne Cedex  
France

Mark Buchanan

Gerard Cagney  
Room F067  
Conway Institute  
University College Dublin  
Belfield  
Dublin 4  
Ireland

G. Caldarelli  
Centro SMC, INFN-CNR  
Dipartimento di Fisica  
Università 'Sapienza'  
Piazzale Aldo Moro 5  
Rome  
00185  
Italy

Elisa Calistri  
Centro Interdipartimentale per lo studio  
delle Dinamiche Complesse (CSDC)  
Università di Firenze  
Via G. Sansone 1  
I-50022 Sesto Fiorentino  
Firenze  
Italy

P. De Los Rios  
Laboratoire de Biophysique  
Statistique, ITP SB EPFL  
Lausanne  
1015  
Switzerland

Thomas M. A. Fink  
Institute M. Curie  
5 Rue d'Ulm 75005  
Paris  
France

A. Gabrielli  
Centro SMC, INFN-CNR  
Dipartimento di Fisica  
Università 'Sapienza'  
Piazzale Aldo Moro 5  
Rome  
00185  
Italy

D. Garlaschelli  
Dipartimento di Fisica  
Università di Siena  
Via Roma 43 Siena  
Italy

S. C. Janga  
MRC-Laboratory of Molecular Biology  
University of Cambridge  
Hills Road  
Cambridge  
CB2 0QH  
UK

Pietro Lió  
Computer Laboratory  
University of Cambridge  
Cambridge  
UK

Seesandra V. Rajagopala  
J Craig Venter Institute (JCVI),  
formerly The Institute of Genomic  
Research (TIGR)  
9704 Medical Center Drive  
Rockville  
MD 20850  
USA

Francesco Rao  
Centro Studi e Ricerche e Museo della  
Fisica 'E. Fermi'  
Compendio Viminale  
Rome  
Italy

Erzsébet Ravasz Regan  
Beth Israel Deaconess Medical Center  
Harvard University  
Boston  
USA

Gian Paolo Rossini  
Dipartimento di Scienze Biomediche  
Università di Modena e Reggio Emilia  
Via G. Campi 287  
Modena  
I-41100  
Italy

Daniel Segrè  
Bioinformatics program  
Boston University  
44 Cummington St.  
Boston, MA 02215  
USA

Björn Titz  
Crump Institute for Molecular Imaging  
Graeber Lab  
University of California, Los Angeles  
570 Westwood Plaza  
Los Angeles  
CA 90095-1770  
USA

Peter Uetz  
The J Craig Venter Institute (JCVI),  
formerly The Institute of Genomic  
Research (TIGR)  
9712 Medical Center Drive  
Rockville  
MD 20850  
USA

Michele Vendruscolo  
Department of Chemistry  
University of Cambridge  
Lensfield Road  
Cambridge  
CB2 1EW  
UK

# Introduction

M. BUCHANAN, G. CALDARELLI, P. DE LOS RIOS, F. RAO  
AND M. VENDRUSCOLO

Biologists now have access to a virtually complete map of all the genes in the human genome, and in the genomes of many other species. They are aggressively assembling a similarly detailed knowledge of the proteome, the full collection of proteins encoded by those genes, and the transcriptome, the diverse set of mRNA molecules that serve as templates for protein manufacture. We increasingly know the “parts list” of molecular biology. Yet we still lack a deep understanding of how all these parts work together to support the complex and coherent activity of the living cell; how cells and organisms manage the concurrent tasks of production and re-production, signalling and regulation, in fluctuating and often hostile environments.

Building a more holistic understanding of cell biology is the aim of the new discipline of systems biology, which views the living cell as a network of interacting processes and gives concrete form to the vision of François Jacob, one of the pioneers in the study of genetic regulatory mechanisms, who spoke in the 1960s of the “logic of life.” Put simply, systems biologists regard the cell as a vastly complex biological “circuit board,” which orchestrates diverse components and modules to achieve robust, reliable and predictable operation. Systems biology suggests that the mechanisms of cell biology can be related to the information sciences, to ideas about information flow and processing in de-centralized networks.

This view, of course, has long been implicit in the study of cell signalling and other key pathways of molecular bio-chemistry. It has long been clear that such pathways do not act through simple sequential action in chain-like reaction paths, but as a rule exhibit a much richer dynamics involving multiple pathways working in parallel, with interactions passing between, and feeding backward as well as forward. Yet with rapidly advancing technology, and new theoretical tools coming from physics, engineering and mathematics, systems biology is beginning to reach beyond the recognition that such systems exist to elucidate specific quantitative

regularities in biological information flows, often involving operations such as feedback, synchronization, amplification and error correction that are familiar to engineers.

Significantly, the potential power of this perspective is being multiplied by an explosion of recent work in physics and mathematics, both theoretical and empirical, showing that many of the world's complex networks have hidden structural regularities of great importance. These networks – ranging from social networks and food webs to the Internet, and including genetic regulatory or metabolic networks within the cell – have a surprisingly universal character, and can be fruitfully described with a unified conceptual language. This work has infused science with new analytical measures such as betweenness centrality, network dimension, degree distribution and so on, which reflect local or global network properties. These measures have been found to have direct relevance to a network's stability or resilience, information processing efficiency, or dynamical richness, offering hope that the network perspective will prove invaluable for understanding networks in cellular biology.

*Networks in Cell Biology* aims to fill a yawning gap in the modern literature of systems biology. While a number of excellent texts at the graduate level cover recent developments in cell biology, and others offer timely introductions to recent advances in complex network science, no single volume yet focuses explicitly on advances in our understanding of cellular networks. The present book offers an up-to-date snapshot of such work, aiming for a balance between empirical studies and theory, which are natural complements in furthering knowledge in almost any area of modern biology. What do we know about the qualitative and quantitative structure and dynamics of genetic regulatory networks, or the network biochemistry underlying cellular metabolism? What do empirical studies tell us about the large-scale architecture of protein–protein interactions, and what are the key weaknesses and gaps in such data? What are the prospects for building realistic dynamical models for metabolic or regulatory processes, possibly with predictive capability?

This volume covers these and other topics of current research in systems biology, with an emphasis on recent advances in complex networks science, and on how the massive data now available in biology can help test and inform such theories in their application. With contributions from key leaders in both network theory and modern cell biology, this volume makes available in one publication the naturally diverse range of studies supporting a unified view of biological networks that is likely to be increasingly important in the future.

The contributions also illustrate a timeless truism of science – that scientific progress is not linear or predictable, and that it is often a new idea emerging from

an unexpected direction – a technique for gathering data of a new kind, or a small shift in perspective – that transforms and renews a field by framing old problems in new ways. The new idea in this case is an old one, the basic perspective of network science, but updated and vastly furthered by modern data and powerful concepts and techniques coming from statistical physics.

# Network views of the cell

PAOLO DE LOS RIOS AND MICHELE VENDRUSCOLO

## 1.1 The network hypothesis

A cell is an enormously complex entity made up by myriad interacting molecular components that perform the biochemical reactions that maintain life. This book is about the network hypothesis, according to which it is possible to describe a cell through the set of interconnections between its component molecules. Hence, it becomes convenient to focus on these interactions rather than on the molecules themselves to describe the functioning of the cell.

The central dogma in molecular biology describes the way in which a cell processes the information required to produce the molecules necessary to sustain its existence and reproduction. It is also becoming increasingly clear, however, that in order to establish a more complete description of the manner in which a cell works we require a deeper understanding of the manner in which the sets of interconnections between these molecules are defining the identity of the cell itself. It is therefore important to investigate whether the genetic makeup of an organism does not only specify the rules for generating proteins, but also the way in which these proteins interact among themselves and with the other molecules in a cell.

**Complex networks** Networks are a way to represent an ensemble of objects together with their relations. Objects are described by means of vertices (sometimes also called nodes) and their relations by edges (sometimes also called links or connections) connecting them, which can be weighted to reflect their strength. A network is thus entirely characterized by the set of its connections, not by the way in which it is drawn. The actual distance between two vertices is given by the minimum number of edges that can be found between them. A key quantity to characterize networks is the degree: it is defined as the number of edges per vertex. Most networks present in nature have large fluctuations in the degree value. This feature has profound consequences both on the stability of the systems represented by the network and on the dynamics of the processes defined on this structure.



Networks provide a way to organize and regulate efficiently complex systems [12, 52, 88, 89, 98, 153, 436, 583, 636]. In an effective network different parts are linked by reducing at a minimum the number of the interconnections. This feat can be achieved at the cost of a fragility of the structure of connections. If one passage goes wrong, the whole structure falls down. A bit of redundancy is a good compromise for a fail-safe system without too much effort. A trade off between redundancy and cost is probably at the basis of the statistical properties of such objects and explains in part their ubiquitous presence in the various cell activities. From the point of view of the researcher a network is also a powerful method to represent the data in one object, and to enable the quantitative assessment of the fragility or robustness of the system. The first step is to establish the topology, the second is to establish the dynamics – that is how the topology changes with time. The study of these two aspects is at the heart of current research about biological networks and it constitutes the core of the following chapters in this book.

## 1.2 The central dogma and gene regulatory networks

**The central dogma** The central dogma, which was introduced in 1958 by F. Crick [131] and later restated in 1970 [132], defines the way in which information flows between DNA, RNA and proteins. At the most fundamental level, DNA passes information to itself (through replication) and to RNA (through transcription), and RNA passes it to proteins (through translation). We also now know, however, that additional information also follows other routes that taken together link DNA, RNA and proteins in a much interconnected network. This view of the central dogma defines our current core understanding of molecular biology.

The central dogma of molecular biology defines the manner in which the genetic information is exploited to generate proteins. In this process, DNA is first transcribed into messenger RNA (mRNA), which is then translated into proteins (Fig. 1.1a). This description, originally proposed by Francis Crick [131], is at the foundation of molecular biology.

Our understanding of the more subtle aspects and of the fundamental implications of the central dogma is becoming increasingly detailed. While initially the information was thought to flow in one direction from genes to proteins, it is now well established that there are much more articulate interactions between DNA, RNA and proteins. Gene transcription is a highly regulated process, in which a major role is played by transcription factors, which are proteins that bind

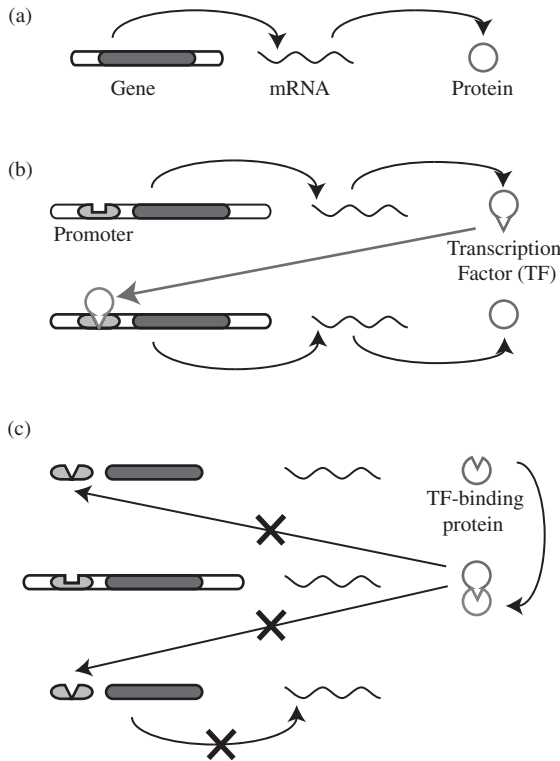


Fig. 1.1. The flow of information between genes and proteins is highly regulated in the cell by a network of interacting molecules. (a) In the simplest view, genes are transcribed into mRNA molecules, which are then translated into proteins. (b) The transcription of genes is controlled by transcription factors that bind to specific promoter regions that precede the genes themselves. (c) The activity of transcription factors can be regulated in a number of manners, including by binding with co-regulatory proteins that block their ability to bind to their corresponding promoters.

to specific promoter regions and control the transcription of the corresponding genes. As transcription factors are gene products themselves a backward flow of information from proteins to DNA is indeed present in the form of a gene-to-protein-to-gene control (Fig. 1.1b). For example, some genes encode proteins that bind to transcription factors, inactivating them. In this case the regulation implies a gene-to-protein-to-protein-to-gene control (Fig. 1.1c).

Furthermore, the cross-specificity between transcription factors and promoter regions is not extreme, allowing for some promiscuity. Indeed the same transcription factor can bind to several promoters and the same promoter can bind to several transcription factors. Moreover, promoter regions are often duplicated so

that several genes might have the same promoters and are regulated by the same transcription factors (Fig. 1.1c).

Our understanding of the mechanism of regulation of gene expression is rapidly increasing. It has been clarified that RNA molecules have a variety of roles in the regulation of protein expression. For example, the recently discovered *small interfering RNA* (siRNA) and *microRNA* (miRNA) [513] molecules bind to the mRNA products of specific genes and inhibit or even enhance their further translation into proteins. This is a gene-to-RNA-to-gene form of control adding further layers of complexity to the way in which gene expression is regulated. It is thus impossible to represent such a rich pattern of relations as simple unidirectional flows. Rather, a network representation is particularly well suited to capture the feedback and feedforward regulation mechanisms that modulate gene and protein expression. Moreover, the network of gene regulation is highly dynamic in order to respond to the changing environment of a cell and to the different requirements through the cell cycle, and should therefore be considered in a time-dependent manner. Chapters 2 and 3 provide a description of the modes of regulation of gene expression, of their actuators and of the theoretical methods in use to reconstruct gene regulatory networks.

### 1.3 Protein–protein interaction networks

**Protein–protein interaction network** A protein–protein interaction network is a graph whose vertices are proteins and edges represent interactions between them, including for example those required to form macromolecular complexes or to establish signaling processes. Given this rather general definition, it is easy to understand why protein–protein interaction networks are becoming a very popular way to represent a variety of functions into the cell. Because of the disparate nature of these interactions, a variety of experimental techniques have been developed to detect them. These methods can be divided into physical (e.g. concentration, immunoprecipitation), library-based (e.g. protein probing, two-hybrid systems) and genetic methods (e.g. synthetic lethal effects) [474].

It is increasingly clear that the majority of proteins do not carry out their functions in isolation but by interacting together in complex manners [506]. There are two main ways to look at protein–protein interaction networks – physical and logical.

Physical interaction networks define interacting protein pairs according to physico-chemical principles. The characterization of this type of network is promoting strong links between structural biology and systems biology, and it is

becoming possible to obtain such networks from experiments in which it is only known that a protein expressed by a given gene can interact with a certain set of other proteins, without the need of knowing the structures, or even the functions, of each of these proteins. The majority of such networks are currently *non-weighted*: only the presence or absence of specific associations are known, but not their strength.

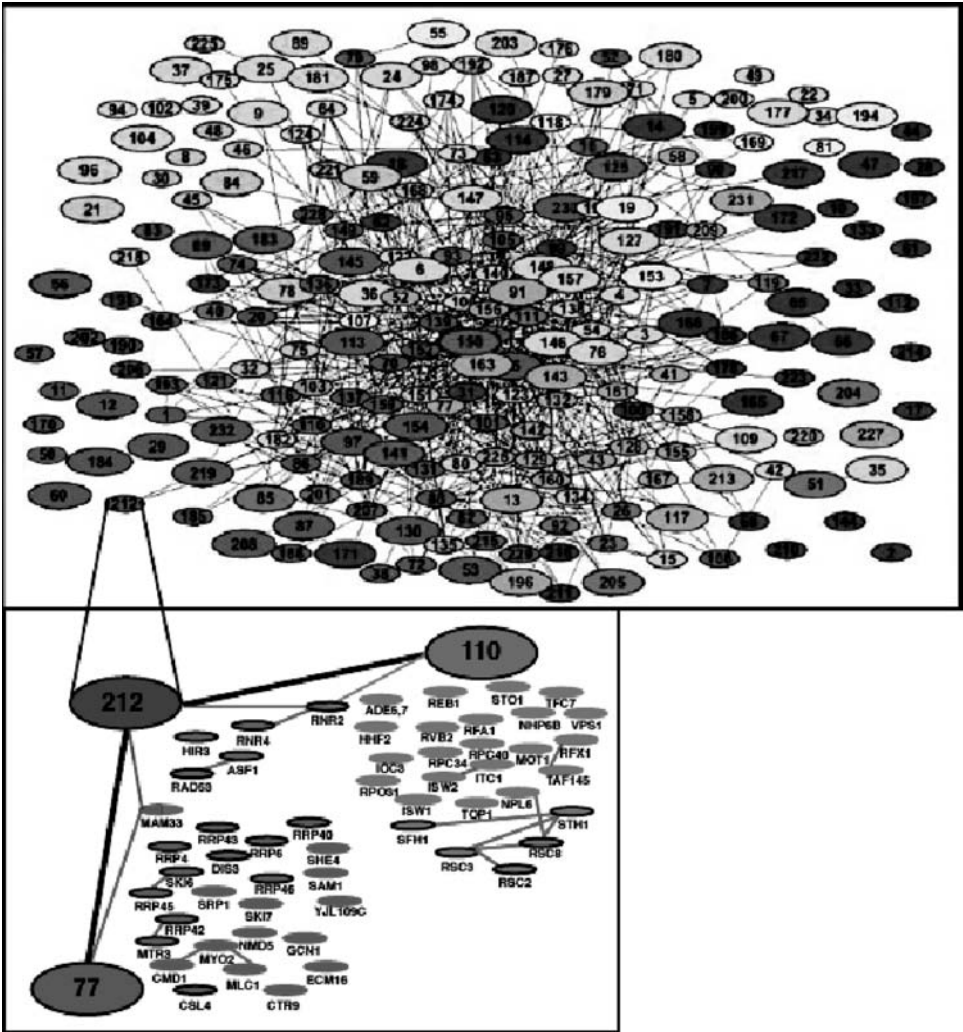


Fig. 1.2. Logical protein interaction network: proteins interact to form complexes, and different complexes may share the same proteins. When this is the case, network edges connect the complexes. Figure from [215].

Logical protein–protein interaction networks come into play once structures beyond the basic dimers become relevant. Indeed, the number of possible multimers grows geometrically and it is currently impossible to predict from basic principles all the possible assemblies. Mass spectrometry techniques allow to identify possible protein complexes, without strictly knowing which proteins are actually in contact with which other within the complex. Thus, these networks provide a glimpse of the logical organization of protein–protein interactions into functional collective units (Fig. 1.2).

Chapters 4 and 5 described how protein–protein interaction networks can be discovered by experiments and by theoretical inference.

## 1.4 Metabolic networks

**Metabolism** Metabolism refers to the ensemble of chemical processes through which living organisms transform resources taken from their environment in the molecules necessary for carrying out cellular functions. Since the products of a chemical reaction are often substrates (i.e. input molecules) of another one, the ensemble of metabolic processes can be conveniently organized as a network. The various chemical compounds present in the reactions are called *metabolites* and reactions are most often catalyzed by specific proteins called *enzymes*. In a metabolic network vertices represent metabolites and edges connect metabolites if they participate in the same reaction.

Simple metabolic processes consist of linear sequences (or pathways) of chemical transformations that take an initial set of molecules and transform them, by leaving by-products on the way, into different ones until the final products are obtained (Fig. 1.3a). A better exploitation of resources can be achieved by using the by-products of a pathway as inputs in other pathways (Fig. 1.3b). Metabolic pathways are indeed highly coupled, an organization that is facilitated by the fact that all the individual reactions take place within a confined space and therefore metabolites are readily exchangeable between different pathways (Fig. 1.4).

Since all reactions acting on the same substrates extract them from the same pool, a competition is permanently present among different pathways. This competition is regulated in a variety of ways, including a co-evolutionary fine-tuning of the enzymatic activities that allows all the reactions not to overexploit the resources to the detriment of the others; also, the different times at which different pathways might be active, thanks to the time-regulation of the process of enzyme expression,

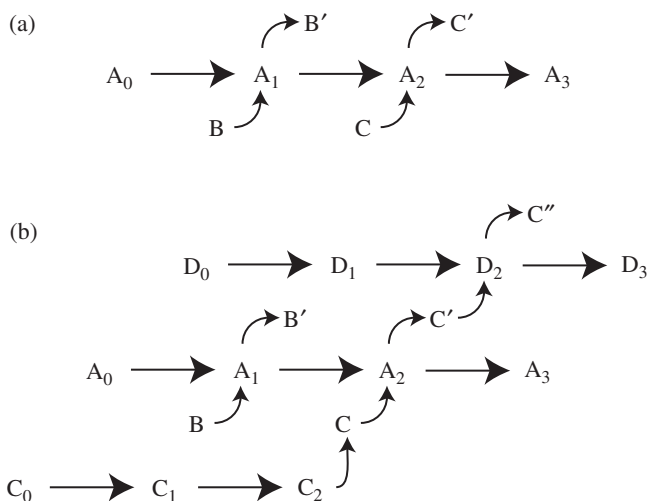


Fig. 1.3. Progressive intertwining of biochemical pathways. (a) A set of molecules  $A_0$  is transformed at first into a second set  $A_1$ , through reactions that involve another set of molecules  $B$ , which in turn transforms into a new set  $B'$ . Then,  $A_1$  is further transformed into a set  $A_2$  through reactions with a set of reagents  $C$ , that is transformed into  $C'$ . After further reactions the end-product  $A_3$  is achieved. (b) The set of molecules  $C$  that is the end-product of a reaction pathway, is coupled to the pathway leading to  $A_3$ , which can coupled to further pathways (e.g. leading from  $D_0$  to  $D_3$ ).

permit different reactions to take place at different times. Part of this fine-tuning is thus linked to the time regulation of gene expression outlined in the previous sections.

Moreover, metabolic networks must be flexible enough to allow even for large variations of their production rates, to cope with different, often significant, changes of the environmental conditions and hence of the needs of the cell. Also, they must be redundant, in order to make sure that the failure or degradation of a pathway is not, to as large a degree as possible, lethal to the cell.

Thus, there is a staggering number of requirements that evolution has had to satisfy while setting up metabolic networks, and it is likely that a full understanding of their rich structures and dynamical features will come only when most of such requirements will have been clarified and included in the theoretical models that we are developing.

Chapters 6 and 7 describe different theoretical techniques to analyze metabolic networks, to unveil their underlying hierarchical structure and to predict their behavior under different environmental conditions.

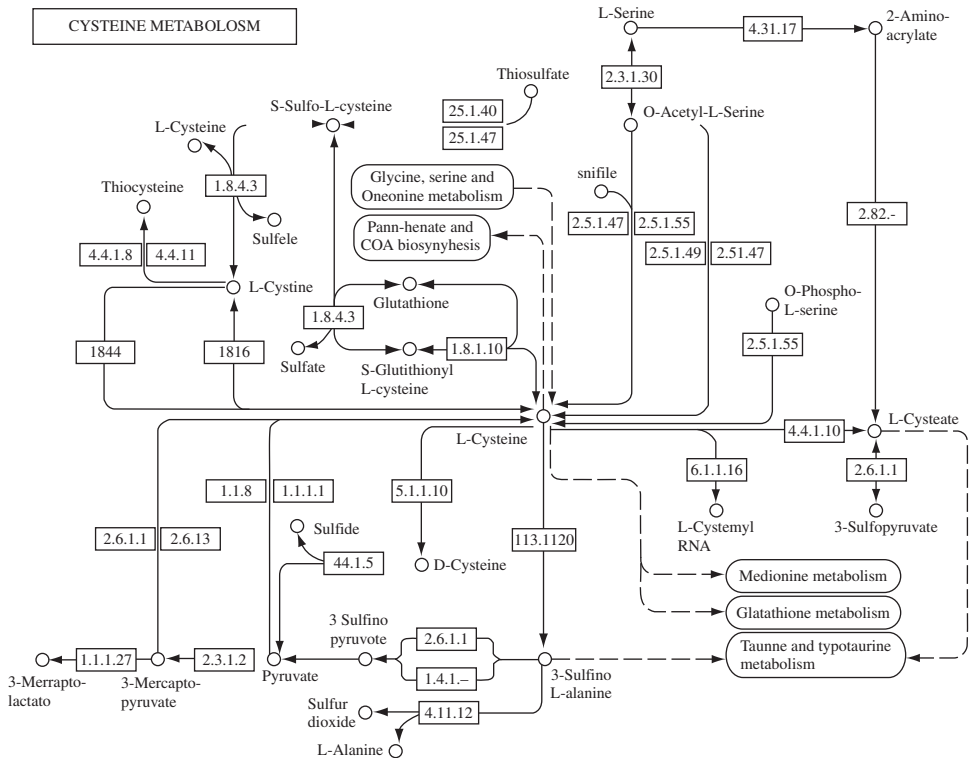


Fig. 1.4. Metabolic network for the synthesis of the amino acid cysteine (from the KEGG database, <http://www.genome.jp/kegg/>).

## 1.5 Signaling networks

**Signaling networks** Signaling networks enable cells to sense changes in their environments and articulate the appropriate responses to them. These networks are made up by collections of interacting signaling pathways, which are cascades of biochemical reactions through which the signals corresponding to the external stimuli are transported to the repository of the genetic code, where they regulate the production of the specific proteins required to orchestrate the overall cellular response.

In order to survive, cells need to respond promptly to the challenges posed by their ever changing environments. They should therefore be able to sense the signals present in the environment and to react to them in the correct way. This response is achieved primarily by means of a series of sensor proteins located on the cell surface, which can alter their behavior upon changes in

pH, temperature or salt concentrations, as well as upon binding to a variety of ligands.

After these signals have reached the cytoplasm, they must be collected and further reported where necessary, typically to genes that encode the proteins needed for the cellular response. For example, in many signaling pathways, sensor proteins are able to activate protein kinases, whose role is to phosphorylate (i.e. attach a phosphate group) other proteins, which as a consequence are activated or deactivated. The change of activity of such phosphorylated proteins ultimately results in the regulation of the expression of the genes responsible for generating the response of the cell to the specific external stimulus. Eventually a host of proteins are produced that not only provide the correct change of cell functionality, but also, down the line, shut off the signal, thus effectively down-regulating their own production once the cellular response is completed.

Chapter 8 is devoted to highlighting the different types of signal transduction in the cell and how they intertwine.

## **1.6 Networked networks and cell functionality**

The networks that we have outlined do not work in isolation but, rather, in a highly concerted manner. Interactions between genes are mediated by RNA and protein molecules, whose building blocks (nucleic acids and amino acids, respectively) are the end-products of metabolic networks. In turn, the reactions present in metabolic networks (e.g. fatty acid synthesis [385]) are almost invariably catalyzed by enzymes that are end-products of gene regulatory networks, which involve a myriad of protein–protein interactions. Moreover, different modes of gene regulations are enacted depending on external and internal conditions (changes of temperature, nutrients, salts, etc.). The transduction of these signals depends on the regulation of protein–protein interactions, such as kinases with membrane proteins and with their targets, of complexes involving transcription factors and of the end-product proteins with the membrane proteins themselves to turn off the signal. Thus, gene expression networks are regulated by signaling networks and transduced via protein–protein interaction networks.

These are just a few examples of the seamless connection between the different classes of biological networks that are crucial to cell functionality. The ultimate goal of the network approach to cell biology is to understand the whole system within a unified framework. Although achieving this objective is still beyond our capabilities, great advances will be possible by approaches that enable different parts of the whole to be individually analyzed.



## **1.7 Concluding remarks**

We have outlined the reasons for which networks are set to become a major paradigm in biology. Indeed the advantages of describing the interactions between the molecular components of a cell in terms of networks are becoming increasingly recognized. This approach enables to illustrate in an efficient manner the hierarchical organization of a cell, which consists of intertwined networks on a wide range of scales, from the interactions between individual metabolites or proteins to the interplay between metabolic and signaling networks.

# Transcriptional regulatory networks

SARATH CHANDRA JANGA AND M. MADAN BABU

## 2.1 Introduction

Complex systems that describe a wide range of interactions in nature and society can be represented as networks. In general terms, such networks are made of nodes, which represent the objects in a system, and connections that link the nodes, which represent interactions between the objects. In mathematical terms, a network is a graph which comprises of vertices and edges (undirected links) or arcs (directed links). Examples of complex networks include the World Wide Webs, social network of acquaintances between individuals, food webs, metabolic networks, transcriptional networks, signaling networks, neuronal networks and several others. Although the study of networks in the form of graphs is one of the fundamental areas of discrete mathematics, much of our understanding about the underlying organizational principles of complex real-world networks has come to light only recently. While traditionally most complex networks have been modeled as random graphs, it is becoming increasingly clear that the topology and evolution of real networks are not random but are governed by robust design principles.

A number of biological systems ranging from physical interaction between biomolecules to neuronal connections can be represented as networks. Perhaps the classic example of a biological network is the network of metabolic pathways, which is a representation of all the enzymatic inter-conversions between small molecules in a cell. In such a network, nodes represent small molecules, which are either substrates or products of an enzymatic reaction, and directed edges represent an enzymatic reaction that converts a substrate into a product. Yet another cellular network which has been the focus of intense study in the last decade is the network of physical interactions between proteins and is usually referred to as the “protein interaction network.” In such a network, nodes represent proteins and two nodes are connected if the proteins are experimentally determined

to physically interact with each other. The focus of this chapter is on another important class of biological networks, referred to as the “transcriptional regulatory network.” The expression of genes in all living systems, i.e. the synthesis of transcripts, can be positively or negatively regulated by the presence of other proteins called transcription factors (TFs). These TFs physically bind to specific sequences i.e. *cis*-regulatory motifs, near target genes (TG) and affect their expression. It is through the transcriptional regulatory network that a cell co-ordinates its response to both external and internal stimuli by controlling the expression of thousands of genes in appropriate amounts under different conditions and time. At an abstract level, all the regulatory interactions linking TFs to the set of transcriptionally controlled target genes (TGs) in an organism can be viewed as a directed graph, in which the TFs and TGs represent the nodes while the regulatory interactions that connect them are represented as directed edges (see Fig. 2.1). The resulting network is typically a complex, hierarchical, multi-layered graph that can be studied at several levels of detail. Though simple transcriptional regulatory networks involving only a few nodes were one of the first dynamical systems for which modeling attempts were carried out [311–313], substantial advance in our understanding of these networks and their modeling on a genomic scale has

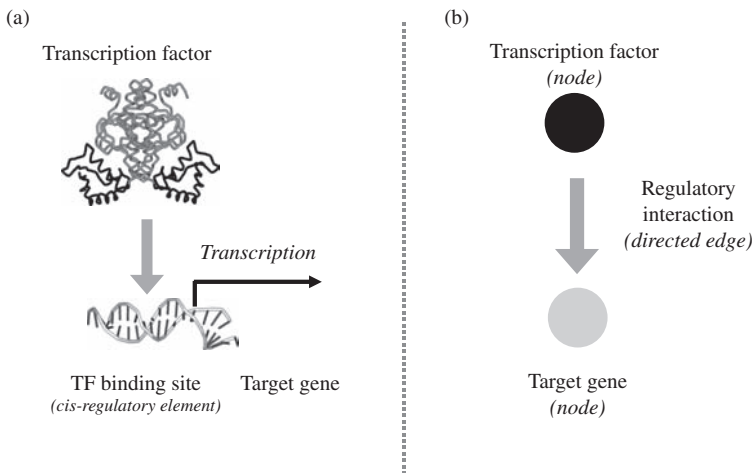


Fig. 2.1. Transcriptional regulatory interaction. (a) Standard view: regulation of gene expression is mediated by transcription factors that bind DNA through their DNA-binding domain. The specific DNA sequence to which they bind is called a TF binding site or *cis*-regulatory element. Binding to the *cis*-regulatory element may either activate or repress the transcription of a nearby target gene (black box). The arrow head represents the transcriptional start site. (b) Network view: transcription factors and target genes are represented as nodes and a directed link connecting a transcription factor to a target gene represents a regulatory interaction between the two.

Table 2.1. *Computer programs, databases and internet based platforms for investigating transcriptional regulatory networks*

<b>Network visualization</b>	<b>Website</b>
Pajek	<a href="http://vlado.fmf.uni-lj.si/pub/networks/pajek">http://vlado.fmf.uni-lj.si/pub/networks/pajek</a>
Cytoscape	<a href="http://www.cytoscape.org/">http://www.cytoscape.org/</a>
Osprey	<a href="http://biodata.mshri.on.ca/osprey/index.html">http://biodata.mshri.on.ca/osprey/index.html</a>
GraphViz	<a href="http://www.graphviz.org/">http://www.graphviz.org/</a>
H3Viewer	<a href="http://graphics.stanford.edu/munzner/h3/">http://graphics.stanford.edu/munzner/h3/</a>
Visant	<a href="http://visant.bu.edu/">http://visant.bu.edu/</a>
Biolayout	<a href="http://cgg.ebi.ac.uk/services/biolayout/">http://cgg.ebi.ac.uk/services/biolayout/</a>
Yed	<a href="http://www.yworks.com/">http://www.yworks.com/</a>
NetMiner	<a href="http://www.netminer.com/">http://www.netminer.com/</a>
<b>Network analysis</b>	<b>Website</b>
NEAT	<a href="http://rsat.ulb.ac.be/rsat/index_neat.html">http://rsat.ulb.ac.be/rsat/index_neat.html</a>
Mfinder	<a href="http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html">http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html</a>
FanMod	<a href="http://www.minet.uni-jena.de/wernicke/motifs/">http://www.minet.uni-jena.de/wernicke/motifs/</a>
Clique finder	<a href="http://topnet.gersteinlab.org/clique/">http://topnet.gersteinlab.org/clique/</a>
MCode	<a href="http://cbio.mskcc.org/bader/software/mcode/index.html">http://cbio.mskcc.org/bader/software/mcode/index.html</a>
Cytoscape	<a href="http://www.cytoscape.org/">http://www.cytoscape.org/</a>
Vanted	<a href="http://vanted.ipk-gatersleben.de/">http://vanted.ipk-gatersleben.de/</a>
Biotapestry	<a href="http://www.biotapestry.org/">http://www.biotapestry.org/</a>
TYNA / Topnet	<a href="http://tyna.gersteinlab.org/tyna/">http://tyna.gersteinlab.org/tyna/</a>
NCT	<a href="http://chianti.ucsd.edu/nct/">http://chianti.ucsd.edu/nct/</a>
Bioconductor	<a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a>
<b>Databases</b>	<b>Website</b>
RegTransBase	<a href="http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=main">http://regtransbase.lbl.gov/cgi-bin/regtransbase?page=main</a>
RegulonDB	<a href="http://regulondb.ccg.unam.mx/">http://regulondb.ccg.unam.mx/</a>
DBTBS	<a href="http://dbtbs.hgc.jp/">http://dbtbs.hgc.jp/</a>
Coryneregnet	<a href="http://www.coryneregnet.de/">http://www.coryneregnet.de/</a>
Prodoric	<a href="http://www.prodoric.de/">http://www.prodoric.de/</a>
Yeast regulatory network	<a href="http://www.mrc-lmb.cam.ac.uk/genomes/madanm/tfcomb/tnet.txt">http://www.mrc-lmb.cam.ac.uk/genomes/madanm/tfcomb/tnet.txt</a>

Table 2.2. *Genome-scale experimental methods to probe protein–DNA interactions*

Method	Description
Chromatin-Immunoprecipitation-Chip experiments (ChIP-chip) Chromatin-Immunoprecipitation-Sequencing experiments (ChIP-seq)	<p>The DNA binding protein is tagged with an epitope and is expressed in a cell. The bound protein is covalently linked to DNA by using an in vivo cross-linking agent such as formaldehyde. After cross-linking, DNA is sheared and the protein–DNA complex is pulled down using an antibody against the tag. Reversal of the cross-link releases the bound DNA, allowing the sequence of the bound fragments to be determined by hybridization to a microarray (ChIP-chip) or by sequencing (ChIP-seq).</p> <p>In ChIP-chip experiments, intergenic regions are spotted on to a microarray chip. Following a chromatin immunoprecipitation step, the bound fragments are reverse cross-linked and hybridized onto the chip. Complementary sequences will bind to specific spots on the chip, thereby providing the exact intergenic region to which the protein was bound [90, 265, 355].</p> <p>In ChIP-seq experiments, the bound fragments are directly sequenced using 454/Solexa/Illumina sequencing technology. The sequences are then computationally mapped back to the genome sequence. Fragments that were bound by the protein will be enriched and hence sequenced several times, providing a direct measure of its enrichment [194, 269, 296].</p>
DNA adenine methyltransferase Identification (DamID)	<p>To overcome any potential non-specific cross-linking of protein to DNA as could happen with ChIP-chip experiments, the DamID technique was introduced. The protein of interest is fused to an <i>E. coli</i> protein, DNA adenine methyltransferase (Dam). Dam methylates the N6 position of the adenine in the sequence GATC, which occurs at reasonably high frequency in any genome (1 site in 256 bases). Upon binding DNA, the Dam protein preferentially methylates adenine in the vicinity of binding. Subsequently, the genomic DNA is digested by the DpnI and DpnII restriction enzymes that cleave within the non-methylated GATC sequence, and remove fragments that are not methylated.</p>