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

Networks in Cell Biology, ed. M. Buchanan, G. Caldarelli, P. De Los Rios, F. Rao and M. Vendruscolo.
Published by Cambridge University Press.

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.

1

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 sytems represented by the network and on the dynamics of the processes defined on this structure.

Networks in Cell Biology, ed. M. Buchanan, G. Caldarelli, P. De Los Rios, F. Rao and M. Vendruscolo.
Published by Cambridge University Press.

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

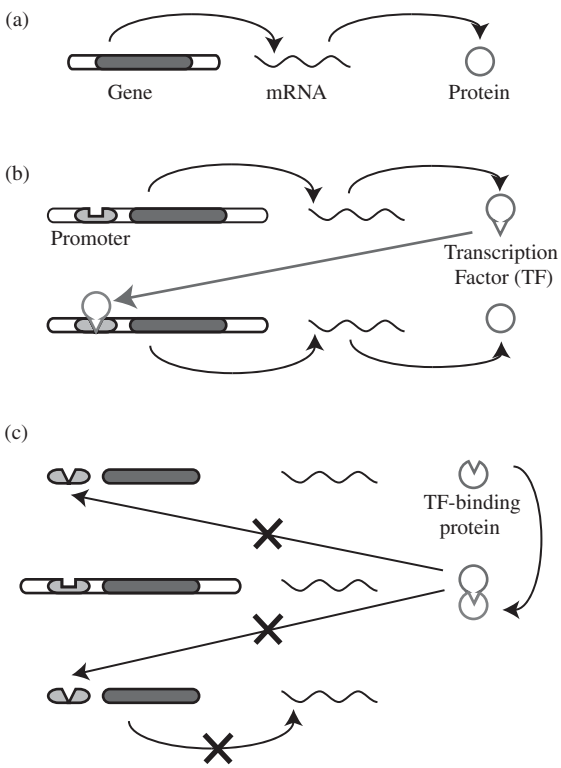

Fig. 1.1. The flow of information between genes and proteins is highly regulated in the cell by a network of interacting molecules. (a) In the simplest view, genes are transcribed into mRNA molecules, which are then translated into proteins. (b) The transcription of genes is controlled by transcription 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.

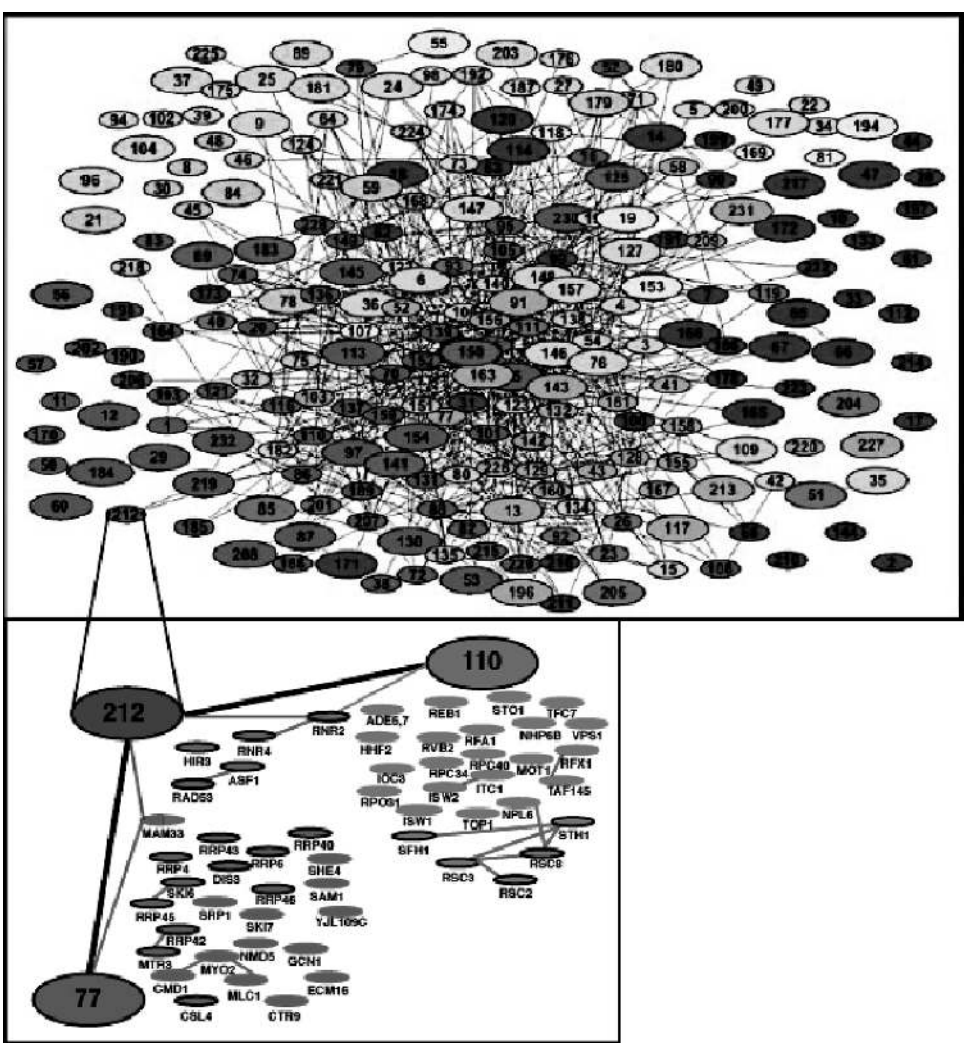

Fig. 1.2. Logical protein interaction network: proteins interact to form complexes, and different complexes may share the same proteins. When this is the case, network edges connect the complexes. Figure from [215].

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,

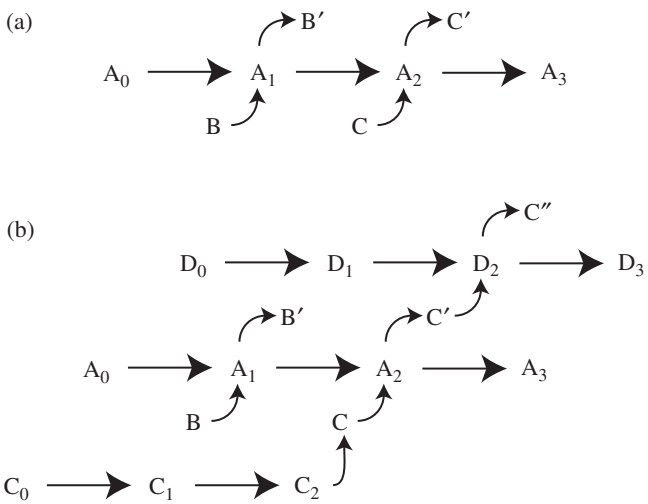

Fig. 1.3. Progressive intertwining of biochemical pathways. (a) A set of molecules A_0 is transformed at first into a second set A_1, through reactions that involve another set of molecules B, which in turn transforms into a new set B'. Then, A_1 is further transformed into a set A_2 through reactions 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.