PART I

NETWORKS IN BIOLOGY

1 An Introduction to Biological Networks

Nuria Planell, Xabier Martinez de Morentin and David Gomez-Cabrero

1.1 Biology Needs to be Analysed Like a System

From basic biology to clinical research, scientists are trying to elucidate the mechanisms underlying the regulation of cells in order to understand their origin, evolution and behaviour in health and disease. Nowadays, we know that the human body comprises a considerable number of different cell types working in coordination. Within a cell, the following framework depicts our understanding of *the biological information* flow from the genome to the phenome: first, the DNA molecules (genomics) are transcribed to mRNA (transcriptomics) and then translated into proteins (proteomics), which can catalyse reactions that act on and give rise to metabolites (metabolomics), glycoproteins and oligosaccharides (glycomics), and various lipids (lipidomics). Finally, these proteins and biomolecules are involved in different metabolic pathways and cellular processes that, in conjunction, dictate the cell behaviour or phenotype [1].

The study of each one of these layers of information (genomics, transcriptomics and proteomics, among others) independently has been extensive and, as a result, there is significant knowledge of the sophisticated machinery that orchestrates the cellular processes. Furthermore, within each layer, many single features (e.g. single genes) have been the target of extensive research, such as the TP53 protein [2–4]. The single-feature analysis derives *partially* from historical technical limitations and from the *belief* that one gene produced a single protein and that one protein had a single function. As a result, there are many *single-gene vs single disease analyses* [2–4]. However, many genes produce several protein isoforms and proteins may have different functions and cellular roles, depending on their environment [5]. Most importantly, many features interact and the 'single-feature' analysis does not allow characterizing such interactions or the behaviours derived from them. Importantly, most cellular functions are organized as highly connected sets of genes and/or proteins and/or metabolites communicating through biochemical and physical interactions. Therefore, *biology needs to move to a holistic view and start to explore all the biological information in an integrated way*: **as a system**. Now, we need *to identify (the best) ways to model biological systems* [6, 7].

One way is to focus on the features and their interactions (whatever the nature of such interactions) and, as a result, a biological system can be depicted as a *network* [8]. In such a *biological* network, the components (nodes) can be genes, proteins or metabolites, among other elements, and the interactions can be physical interactions, biochemical interactions or co-expression, among others.

To illustrate the concept, we will detail an example: a pathogen (for instance, a virulent strain of *Escherichia coli*) infecting our body. When this happens, the immunological response is activated to eradicate the infection and restore a healthy status. At the cellular level, it means that different processes are initiated to produce a pathogen-related response. As a brief description, these processes start with a signal (stimulus) that triggers a sequence of (chemical or physical) signals that are transmitted through the cell, provoking a signal cascade that results in a cellular response. Any process that starts from a particular stimulus and is transformed into a biochemical signal throughout the cell is known as a *signal transduction process* (and these are all good candidates for network modelling).

As a detailed example, we consider one of the signal transduction processes activated as a pathogen-related response, the TLR4 (Toll-like receptor 4) signal transduction pathway. The interaction between the pathogenic molecule and the cellular receptor TLR4 initiates the signal transduction by recruiting intracellular adaptor molecules such as myeloid differentiation factor 88 (MyD88) and TIRF-related adaptor proteins. Depending on the adaptor proteins recruited, two different signal cascades can take place: one that depends on the MyD88 molecule and another which is TRIF-dependent. Following the MyD88-dependent pathway, after the recruitment of adaptor proteins, TNF receptor-associated factor 6 (TRAF6) is activated to interact with the second complex of proteins (TAK1 and TAB2/3). Going forward, mitogen protein kinases (MAPKs; MKK3/6 and MKK4/7) and another complex of proteins (NEMO/IKK complex) are activated, leading to the activation of AP1 (through p38 or c-Jun N-terminal kinase (JNK)) and NF-κB, respectively; all are involved in the transcription control of pro-inflammatory cytokines (IL-6, IL-12, TNF-α, etc.). The MyD88-independent pathway recruits TRIF-dependent adaptor proteins and starts the signal cascade by binding to the IKK-related kinase TBK1 and IKKε, which mediates direct phosphorylation of IRF3 transcription factor. IRF3 will migrate to the cellular nucleus and promote the transcription of IFN-inducible genes [9, 10]. Briefly, from the initial pathogenic stimulus, a signal cascade starts to lead to the production of inflammatory-related cytokines.

In Figure 1.1a (inspired and partially adapted from [10]), the infection process described is depicted, where the proteins or protein complexes are the nodes and the physical or biochemical interactions the edges. Such description can be further summarized into a network, as shown in Figure 1.1b, where *elements* of the information are ignored (e.g. location in the cell or the type of interaction) and only proteins (*nodes*) and interactions (*edges*) are kept. Following both representations, we can identify and follow the signal cascade from the initial stimulus to the final cellular response. The network representation has a mathematical description and notation that will be introduced in the next section (and further discussed in Chapter 2). Finally, in Figure 1.1c we observe that the network can also be stored as a matrix, where rows and columns denote the proteins, and for an entity in the matrix a '1' (dark grey in the figure) denotes an interaction between both proteins.

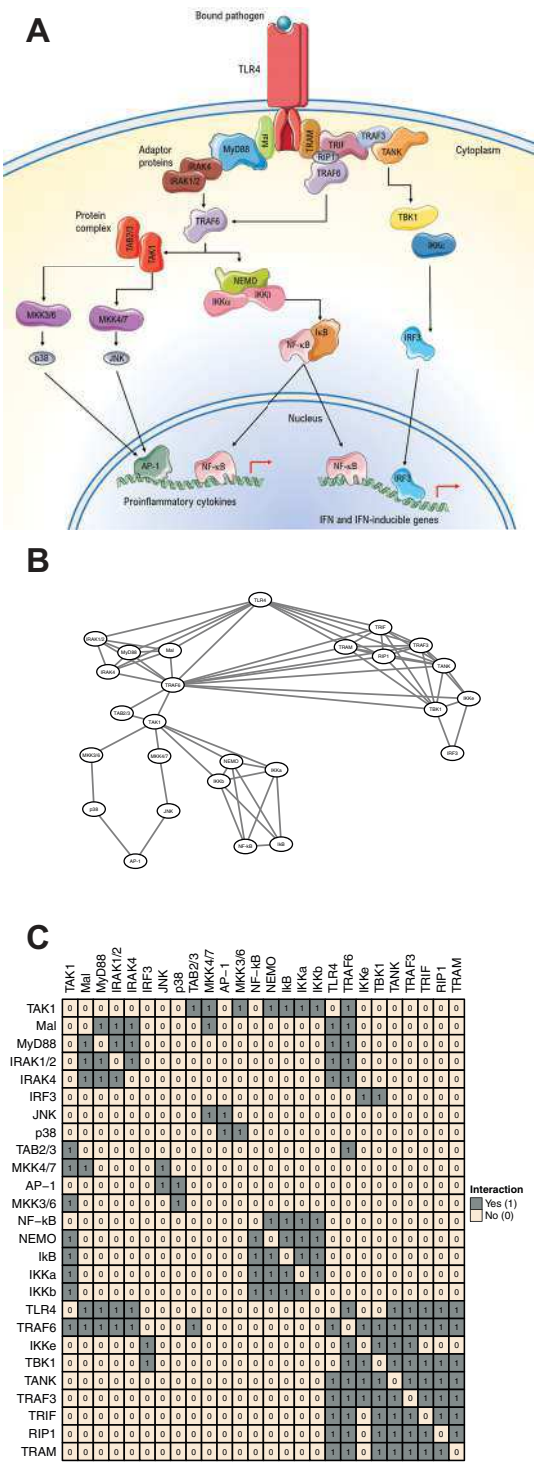

Figure 1.1 Example of a biological network: a signalling transduction network. (a) Biological description. (b) Network description of (a). (c) Contingency matrix description of (a). See text for details of the biological network described [10].

1.2 Introduction to Networks

In the previous section, we established that biological systems can be *modelled* as networks. The role of the modelling here is to provide a holistic description of a system (derived from the biological information) in a way that allows studying characteristics of the systems that cannot be derived from the collection of per-feature characteristics ('stamp collection' [7]). In biological networks, *nodes* can represent any type of biological molecule or even a complex of molecules. *Edges* can represent any type of relationship between a pair of nodes; for example, edges may represent that two molecules are present in the same tissue, are related to the same disease, are part of the same biological process [11] or the same molecular function [12], or similar expression levels [13], among other relationships. In Figure 1.1, a protein–protein interaction network is depicted in which nodes are the proteins and, in some cases, the edges represent known physical interactions [14, 15].

Biological networks can be described as graphs; and, while in the text we will use graph and network interchangeably, we should clarify that network analysis is the study of graphs when they represent relations (symmetric or asymmetric) between discrete objects [16, 17]. Interestingly, the concept of graph theory was initially developed as a tool to solve *mathematical riddles*. The first (and most famous) riddle is the problem of the bridges of Königsberg: the town Königsberg had seven bridges and the problem was to visit all parts of the city while crossing each bridge only once [18]. Euler proved in 1736 that there was no feasible solution [19].

Importantly, around the end of the 1950s, the analysis and generation of random graphs was proposed by Erdös and Rényi [20] and simultaneously by Gilbert [21]. A random graph studies the uniformly random selection of graphs from the set of all possible graphs with N nodes and M edges, with N and M being arbitrary numbers. Interestingly, it was observed that those models were not able to capture a property observed in most 'real-life' networks: small-world properties. A significant small-world property is the short average path length necessary to connect every pair of nodes. Watts and Strogatz proposed a model to generate small-world random graphs [22]. However, those graphs did not generate another 'real-life' network property: 'hubs'. Hubs are (a small number of) nodes with a more extensive than average number of edges to/from other nodes; the property is termed 'scale-free'. Barabási and Albert studied scale-free graph properties [23].

The analysis of random vs non-random graphs is of particular relevance, which we will explore further in a later section, because in biological systems (as well as observed in social networks) the graphs associated are not random graphs as defined by Erdös and Rényi [20]. For instance, there are nodes with an increased number of edges. In gene networks, these nodes are known as 'hubs' or 'master regulators', and they are of interest because they may show an association with specific biological processes.

Computationally, the process of drawing biological systems into networks can be described mathematically by adjacency matrices (see Figure 1.1c). In such a matrix, both columns and edges are the nodes, and every position ($node_i$,$node_j$) may denote the existence/non-existence of an edge as a binary 1/0 (e.g. a protein–protein interaction [24]), or they may specify numerical 'weights' that may be associated with the strength of the relationships. Those weights could be computed as a measure of similarity between the nodes using, for instance, correlation or mutual information

[25], among other measures. It is essential to specify that the selection of the distance measure may shape a biological network differently [26].

1.3 Types of Biological Networks

As previously presented, a biological system can be represented as a biological network, and it may include several groups of coordinated subsystems. From the molecular level up to a whole biological system, one can think in different types of networks: molecular networks, cell-to-cell networks, host–microbiome networks and systems medicine networks. Moreover, each one of these networks can be divided into different subsystems.

Within molecular networks, the most relevant ones are protein–protein interaction (PPI) networks, gene regulatory networks, signal transduction networks, metabolomics or biochemical networks and functional or co-expression networks. The nodes of these networks are genes and/or proteins and/or metabolites, and the edges are physical or biochemical interactions, co-expression patterns, etc. [17, 27]. A schematic representation of the different types of molecular networks is shown in Figure 1.2.

Protein–protein interaction networks are fundamental in biological functions. Protein interactions determine molecular and cellular mechanisms that control healthy and diseased states in organisms. In these networks, nodes represent proteins and edges represent a physical interaction between two proteins. The edges are non-directed, as it cannot be said which protein binds the other; that is, which partner functionally influences the other. Within the example described in Figure 1.1a, several PPI networks can be defined. The interaction between different adaptor molecules and the TLR4 gives a complex structure that is *per se* a PPI.

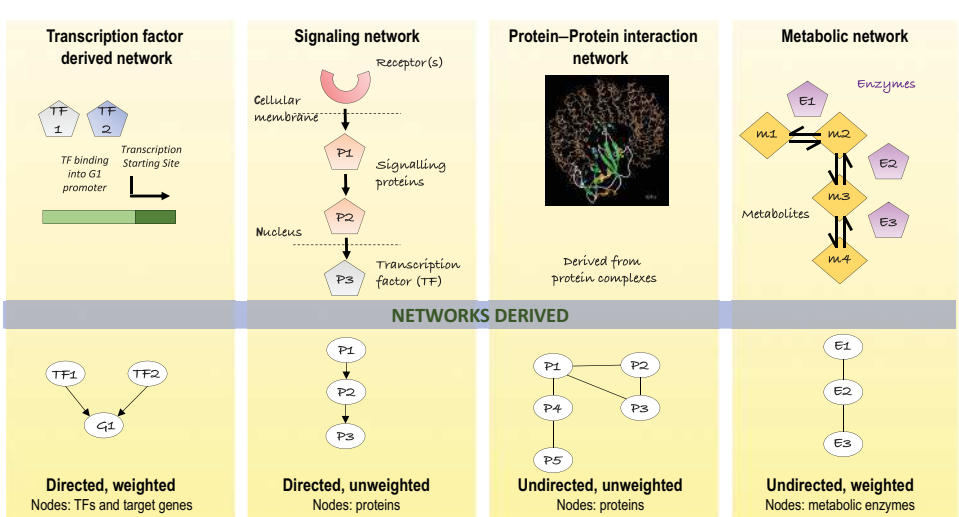

Figure 1.2 Examples of types of biological networks. The image from the Protein-Protein interaction section was created in 2002 by Dcrjsr, and is licensed under the Creative Commons Attribution 3.0 Unported licence.

The structures and dynamics of protein networks are disturbed in complex diseases such as cancer [24] and autoimmune disorders. Therefore, such networks facilitate the understanding of these mechanisms in both pathogenic or physiologic scenarios and can be translated into effective diagnostic and therapeutic strategies [28].

To generate PPI networks, besides the various experimental methods, a variety of large biological databases that collect and organize PPI information are available, most of which are organism-specific. Among them are the Yeast Proteome Database (YPD) [29], the Munich Information Center for Protein Sequences (MIPS) [30], the Molecular Interactions (MINT) database [31], the IntAct database [32], the Database of Interacting Proteins (DIP) [33], the Biomolecular Interaction Network Database (BIND) [34], the BioGRID database [35], the Human Protein Reference Database (HPRD) [36], the HPID [37] and the DroID for *Drosophila* [38]. Additionally, well-documented services based on text-mining analysis provide relevant resources, including the Stitch and String databases [39, 40].

Gene regulatory networks give information concerning the control of gene expression in cells. Nodes are either a transcription factor or a putative DNA regulatory element, and directed edges represent the physical binding of transcription factors to such regulatory elements. Edges are directed: *incoming* (transcription factor binds a regulatory DNA element) or *outgoing* (regulatory DNA element bound by a transcription factor). In addition to transcription factor activities, overall gene transcript levels are also regulated post-transcriptionally by microRNAs (miRNAs), short noncoding RNAs that bind to complementary cis-regulatory RNA sequences usually located in 30 untranslated regions (UTRs) of target mRNAs. Then, edges can also be denoted as *incoming* (miRNA binds a 30UTR element) or *outgoing* (30UTR element bound by an miRNA).

These networks use a directed graph representation to model the way proteins and other biological molecules are involved in gene expression, and they aim to describe the order of the events that take place in different stages of the process. Following the example in Figure 1.1a, the associated (not in the figure) regulatory network of the activated transcription factors AP1 and NF-κB could be detailed.

To generate this regulatory networks, protein–DNA interaction data is collected in databases like JASPAR [41], TRANSFAC [42] or B-cell Interactome (BCI) [43], while post-translational modification can be found in databases like Phospho.ELM [44], NetPhorest [45] or PHOSIDA [46].

Signal transduction networks connect receptors and many different cellular machines. Such networks not only receive and transmit signals, but also process information. To represent the series of interactions between the different biological entities (nodes) such as proteins, chemicals or macromolecules and to investigate how signal transmission is performed either from the outside to the inside of the cell or within the cell, multi-edged directed graphs are used. One example of these signal cascades is shown in Figure 1.1a. Given a pathogenic stimulus, a signal is transmitted through the cell to give a response. Depending on the cellular circumstances (environmental parameters), different responses can be triggered; in that way, the environment could trigger for a MyD88-dependent or -independent response in the case of the TLR4 signalling pathway. Some sources of information regarding signal transduction pathways are the MiST [47] and TRANSPATH [48] databases.

Metabolomics or biochemical networks describe a series of chemical reactions occurring within a cell at different time points. The enzymes play the primary role within a metabolic network since they are the main determinants in catalysing biochemical reactions. Often, enzymes are dependent on other cofactors, such as vitamins, for proper functioning.

In graph representation of metabolic networks, nodes are metabolites and edges are either the enzymes that catalyse these reactions or the reactions that convert one metabolite into another. Edges can be directed or undirected, depending on the reversibility of a given reaction. Among the several databases holding information about biochemical networks, some of the most popular are the Kyoto Encyclopedia of Genes and Genomes (KEGG) [49], EcoCyc [50], BioCyc [51] and metaTIGER [52].

Functional networks are gene co-expression networks. The reasoning used to define this type of network is that associated proteins are more likely to be encoded by genes with similar transcription profiles [53, 54]. In these networks, nodes represent genes and edges link pairs of genes that show correlated co-expression above a set threshold based on an association measure such as the Pearson correlation coefficient or mutual information [55]. In the example shown in Figure 1.1a, the set of genes whose transcription is regulated by NF-κB and AP1, such as IL-6, IL-12, IL-1 and TNF-α, may show statistically significant correlation because they are involved in the same biological process [56].

Beyond molecules, **cell–cell communication (CCC) networks** can also be defined. This kinds of networks describe the cross-talk between cells. In those networks, nodes are different cell types and the edges are receptor–ligand interactions. A CCC network is a directional bipartite graph that is usually constructed based on the differential over-expression of ligand and receptor genes of the cell types of interest [57].

Given the complex system that defines a whole organism and the functional interdependencies between the molecular components shown in a human cell, we observe that most diseases are rarely a consequence of an abnormality in a single gene. Instead, the disease phenotype reflects the perturbations of a complex intracellular network. The identification of these perturbed networks defined as disease modules can allow the identification of molecular relationships between apparently distinct pathologic phenotypes. These disease connections can be presented as a **disease network**, where nodes are disease and diseases are connected if they share one or several disease-associated genes or if they are both associated with enzymes that catalyse adjacent reactions. In metabolic diseases, links induced by shared metabolic pathways are expected to be more relevant than links based on shared genes [58]. To construct this kind of network, available resources are the gene–disease associations collected in the OMIM [59], KEGG [60] and BiGG [61] database.

Other approaches are emerging within systems medicine, including **drug–target networks** and **drug–drug networks**. Both drug–target and drug–drug networks will help in new drug development as they are implicated in drug discovery and prediction of adverse effects [62, 63]. Those types of networks are also described in Chapter 9.

Finally, **microbiome–host networks** can also be defined. The role of the microbiome in human health and disease has received greater interest during recent years as the microbiome is involved in metabolism, physiology, nutrition and different immunological functions. For more in-depth information on microbiome and host–microbiome networks, see Chapter 11.

In summary, several types of networks can be defined in biology in order to explain and simplify complex systems. However, these approaches are restricted to the amount of information known; a vast amount of interactions is thought to be unknown. Consequently, biological networks should be considered as a dynamic field that will evolve over time, depending on knowledge generation and curation.

1.4 Mathematical Properties of Biological Networks

In biological networks, as well as in social networks, the distribution of the number of edges incident upon a node – denoted as *node degree centrality* measure – follows a power law distribution, $P(k) = Ak^{2-y}$ [64], that is not observed in random graphs. As a result of the power-law distribution there will be high diversity of node degrees; this characteristic is known as **scale-free** [23]. A second property is the **small world** [22], which denotes that the 'shortest path' (or the collection of nodes) needed to communicate a pair of nodes is reduced compared to random networks.

An additional property of interest in networks is connectivity, which estimates (and identifies) the minimum number of edges (or nodes) required to separate nodes into isolated subgraphs. Isolated subgraphs are groups of nodes that cannot describe a path connecting them. In Figure 1.1a, the elimination of edges (p38,AP-1), (JNK,AP-1) and (NF-κKB,NF-κB) would generate two subgraphs.

There are also measures of interest that define the relevance of a node, such as centrality measurements. Beyond node degree centrality, *betweenness centrality* quantifies the number of times a node appears in shortest paths between pairs of nodes or *closeness centrality* quantifies the average length of the paths between the node of interest and any other node, among other measures [8]. These properties are described and discussed in Chapter 3.

1.5 Storing and Visualizing Networks

Networks are a useful tool for modelling and studying most biological systems. While the mathematical tools for their analysis are relevant, the storage and visualization of networks are also relevant because they provide powerful exploratory tools.

For storing and communication, the Systems Biology Markup Language (SBML) [65] provides a representation format based on XML, which allows the communication and storage of computational models of biological processes. It's an open-source framework and nowadays is the standard for representing computational models in systems biology.

For visualization, there are many tools available, among the most popular being Cytoscape [66] and Gephi [67]. Both tools provide methods for visualization, but also network analysis (including the estimation of centrality measures) or interfaces with programming languages such as R. Importantly, network visualization is a complex problem by itself, because it requires describing in two dimensions a set of features and their connections. There are several methodologies available for the projection of networks in two dimensions (named layout). Several examples of the network shown in Figure 1.1 are shown in Figure 1.3.