Complex Networks in Genomics and Proteomics

Ricard V. Solé and Romualdo Pastor-Satorras

WILEY-VCH Verlag Berlin GmbH
May 29, 2002

1 Complex Networks in Genomics and Proteomics

Ricard V. Solé and Romualdo Pastor-Satorras

[1]ICREA-Complex Systems Lab, Universitat Pompeu Fabra-IMIM
Dr Aiguader 80, 08003 Barcelona, Spain
[2]Departament de Fisica i Enginyeria Nuclear
Universitat Politecnica de Catalunya
Campus Nord B4, 08034 Barcelona, Spain

1.1 Introduction

Complex multicellular organisms contain large genomes in which each structural gene is associated with at least one regulatory element and each regulatory element integrates the activity of at least two other genes. The nature of such regulation started to be understood from the analysis of small prokaryotic regulation subsystems and the current picture indicates that the webs that shape cellular behavior are very complex. Actually, integration of extracellular signals often involves the crosstalk between signal cascades that has been suggested to share some common traits with neural networks [1]. In a related context, detailed analyses of subsets of interacting genes reveal that cell biology is highly modular [2]. Here "modules" are made up of many species of interacting molecules and the functional relevance of these subnets is highlighted by the observation that they are conserved through evolution.

In many cases, proteins composed by multiple subunits behave as switch-like elements that can flip, for example, from an active to an inactive state and back. The switching behavior of these complexes, together with the underlying information processing that takes place at the network level, allows for a computational description of intracellular signaling. In this context, one might consider some key features of standard computational systems that should apply here. One particularly important aspect is the resilience of the signaling network under different sources of perturbation. The analysis of mutational robustness in different organisms revealed an extraordinary level of homeostasis: in many cases the total suppression of a given gene in a given organism leads to a small phenotypic effect or even to no effect at all [3, 4].

Following the analogy with engineered systems, the immediate explanation for such robustness would come from the presence of a high degree of redundancy. Under mutation, additional copies of a given gene might compensate the failure of the other copy. However, the analysis of redundancy in genome data indicates that redundant genes are rapidly lost and that redundancy is not the leading mechanism responsible for mutational robustness [4].

The origins of robustness against mutations is particularly well highlighted by the analysis of genome-wide scale data of the budding yeast *Saccharomyces cerevisiae* [4]. The main conclusion of this study is that the major cause of robustness comes from the interactions among unrelated genes. This mechanism would be illustrated by the following example: given a

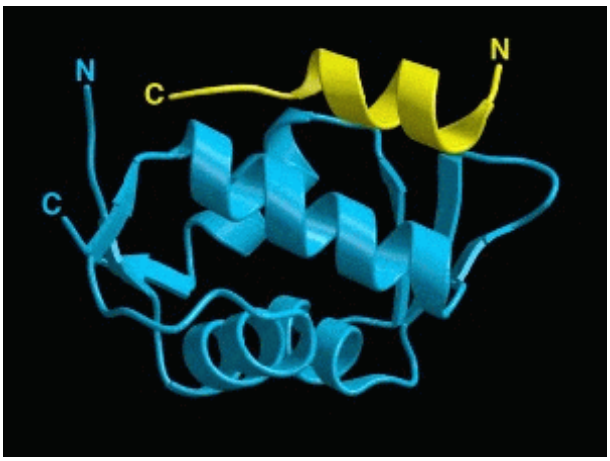

Figure 1.1: The domain of molecular interaction among p53 and MDM2 is shown in this 3D reconstruction [5]. MDM2 (here in cyan) binds a specific domain of p53 in a region (here shown in yellow) important for the interaction of p53 with components of the transcription machinery.

metabolic network, completely unrelated enzymes can catalyse different reactions but contribute to a pathway whose goal is to sustain an optimal flux of metabolites. Under these conditions, mutations in genes encoding those enzymes will have little or mild effects. Additionally, it is interesting to see that many examples of experimental biotechnology manipulations involving the tinkering of one or two genes fail to reach the expected goals: very often, counterintuitive outcomes are obtained.

On the other hand, mutations involving some key genes can have very important consequences. This is the case, in particular, of the p53 tumor suppressor gene, Figure 1.1, which is known to play a critical role in genome stability and integrates many different signals related to cell-cycle or apoptosis (cell death) [6]. This and other tumor-suppressor genes prevent cell proliferation (thus keeping cell numbers under control) but can also promote apoptosis. The example of p53 is particularly important because it is mutated or there is a functional defect in the p53 pathway in approximately half of human cancers. The p53 network (partially shown in Figure 1.2) is quite well-known in mammals and involves genes that control, for example, apoptosis, the development of blood vessels, or cell differentiation. The core of this net is defined by the feedback loop existing between p53 and its negative regulator, the MDM2 oncoprotein. In invertebrates (such as *Drosophila*) homologues of p53 are known to be active throughout early development [7].

The fact that many mutations have little or no effect seems to be consistent with the presence of genes that either cannot propagate their failure or whose function can be replaced by other parts of the net. The presence of some genes that integrate multiple signals and can trigger widespread changes under their failure shows that the underlying network includes some highly-connected hubs. It seems to be a compromise between integration and homeostasis that should be observable when looking at the map of interactions within the cellular net.

Although a complete description of cellular networks would require the explicit consider-

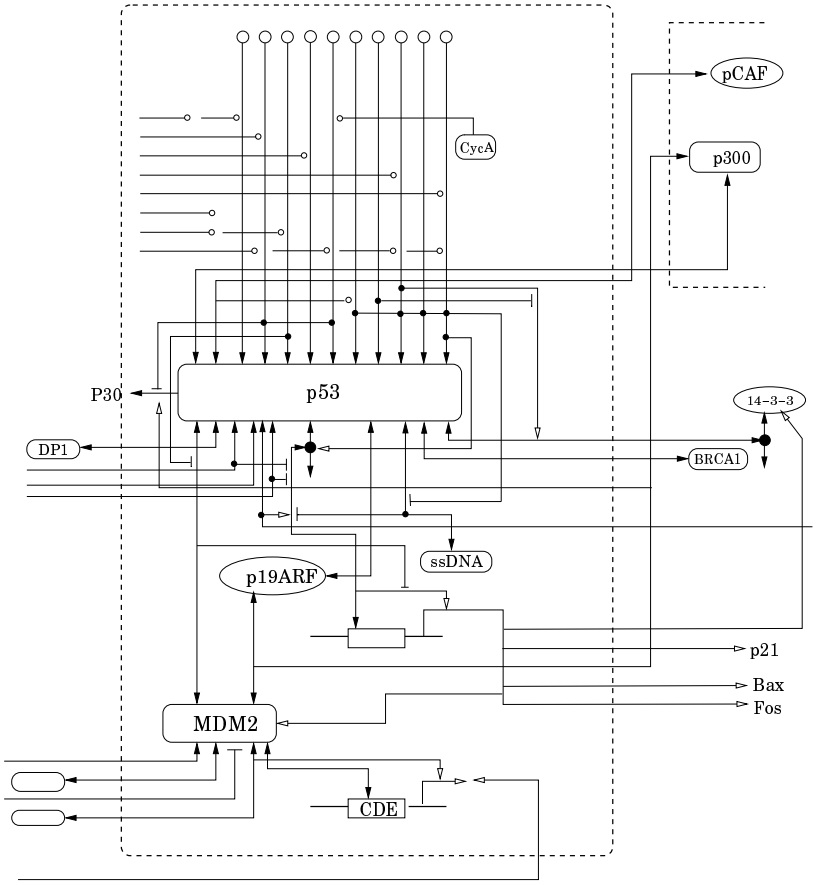

Figure 1.2: Schematic architecture of the p53 network. The p53 node integrates information from very different parts of the system. Only part of the cell circuitry is shown here. For a detailed presentation, see Ref. [8].

ation of dynamics, topological approaches—in which only the static architecture of the net is considered—are often successful in providing insight into biological complexity. This is the case, for example, of some models of ecological networks: in spite that populations fluctuate in time and changes in biomass or productivity take place at different scales, some of the fundamental regularities exhibited by food webs can be fairly well explained by means of static approaches [9]. Besides, the comparison of a wide range of complex networks (both natural and artificial) reveals that strong regularities are shared by them, in spite that their underlying components, the nature of their interactions, and their time scales are very different. In this chapter the topological patterns displayed by these networks will be explored. As will be shown, the compromise between stability and integration can be made explicit by looking at the large-scale organization of cellular networks.

1.2 Cellular networks

The molecular basis of genetic control in cells, particularly in eukaryotic cells (i. e. cells with nucleus) is one of the most basic active areas of molecular cell biology. Of particular interest is the understanding of the regulation mechanisms involved in the development of multicellular organisms. In most well-known case studies, such as in the fruit fly *Drosophila melanogaster*, it has been shown that regulation among the genes that control early development (such as fushi tarazu, Figure 1.3) takes place at the transcription level [10]. The web of interactions can be very complex, and an example of a sub-web of the genetic net associated to *Drosophila* early development is shown in Figure 1.3(b). Mutations in genes associated to early stages of development have typically a strong effect and sometimes, as it occurs with the so-called homeotic genes [11], they result in important morphological changes.

Models of gene regulation have a long history in theoretical biology [12, 13]. The discovery of the mechanisms of transcription regulation in the Lac operon of *E. coli* was followed by the formulation of some simple mathematical models [14]. Inspired in early models of neural networks, a standard formulation of gene regulation can be introduced by means of a dynamical system:

$$\frac{dg_i(t)}{dt} = \Phi_\mu^i[\mathbf{g}] - \gamma_i g_i, \qquad i = 1, ..., n, \tag{1.1}$$

where a set of n different genes is defined. Here $\mathbf{g} = (g_1, ..., g_n)$ gives the activity state of each gene. Degradation is introduced by the last term $\gamma_i g_i$. The function $\Phi_\mu^i[\mathbf{g}]$ introduce the nature and extent of the interactions among components. An example of such type of model is:

$$\frac{dg_i(t)}{dt} = \Phi_\mu^i \left(\sum_{j=1}^{n} W_{ij} g_j(t) - \theta_i \right) - \gamma_i g_i(t), \tag{1.2}$$

where $\Phi_\mu^i(x)$ is a sigmoidal function of the local field $h_i = \sum_j W_{ij} g_j$, θ_i is a threshold, and the weights W_{ij} give the sign and strength of the gene-gene interactions. Usually the set $\mathbf{W} = \{W_{ij}\}$ is generated from a given distribution $\rho(W)$ that is assumed to be symmetric and with zero mean. This type of net can display a huge variety of dynamical patterns, including oscillations and chaos [15]. But the really interesting behavior (see below) comes from the statistical properties derived from the presence of phase transitions [16] when the connectivity is tuned.

Why to consider this type of mathematical approximations? Some attempts of building large-scale models of cellular nets based on near-realistic descriptions have failed to reproduce the whole spectrum of dynamical patterns displayed by even simple controlled systems. On the other hand, some key questions can find powerful answers in the generic properties exhibited by simple representations of real nets [16]. As an example, a striking feature of multicellular diversity is the surprisingly small repertoire of cell types, given the potentially astronomic diversity of cell states that would be obtained from the combinatorics of gene states [17]. Assuming that the number of genes in a multicellular organism is $N \approx 10^4$, 2^N different possible states are available. Yet, if cell types are considered as indicators of gene expression states, only $200 - 300$ states are actually realized.

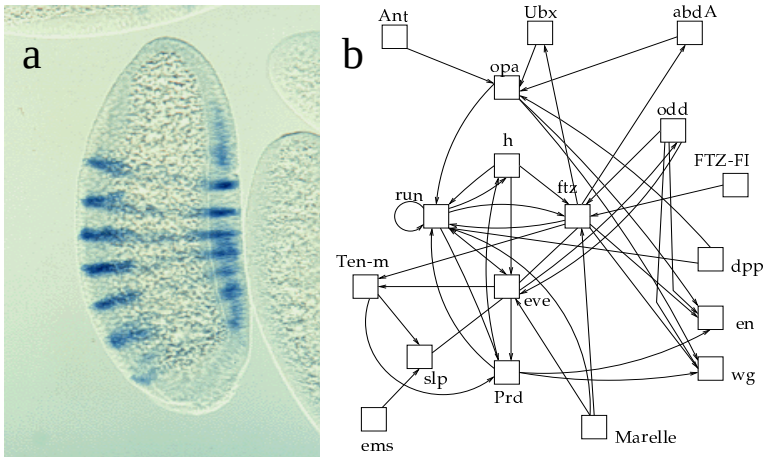

Figure 1.3: (a) Spatial pattern of activity of a given gene involved in *Drosophila* development (the so-called fushi tarazu gene (FTZ); see also Figure 1.2). The darker areas correspond to higher levels of activity of FTZ, indicating what cells are expressing it. Cell-to-cell interactions generate this set of stripes with a characteristic length. In (b) an example of a real gene network is shown. It includes some part (i.e. a directed subgraph) of the genetic net involved in the development of *Drosophila* fly. The names of the genes involved are indicated, such as FTZ=fushi tarazu. Only the connections are shown, not their sign.

In this section we will summarize some key features of this type of dynamical systems by considering the richness of their attractors when low-dimensional nets are used. Afterwards, the general scenario involving a large number of genes (i.e. large networks) will be considered.

1.2.1 Two-gene networks

The minimal number of genes needed in order to obtain a rich spectrum of behavioral patterns is given by two elements in interaction, although single-gene models with the appropriate nonlinearities can also display complex dynamic behavior [18]. Two-gene models allow to understand particularly important problems, such as the dynamics of virus-cell interactions in bacteria [19]. An example is the following two-gene system with no self-interaction, described by the equations:

$$\frac{dg_1}{dt} = \delta \frac{W_{21} g_2}{1 + W_{21} g_2} - g_1 \tag{1.3}$$

$$\frac{dg_2}{dt} = \delta \frac{W_{12} g_1}{1 + W_{12} g_1} - g_2. \tag{1.4}$$

The fixed points are easily found; together with the trivial fixed point, $P_0^* = (0, 0)$ we get a second nontrivial point $P_1^* = (g_1^*, g_2^*)$ given by:

$$g_1^* = \frac{\alpha}{W_{12} + \Omega} \quad g_2^* = \frac{\alpha}{W_{21} + \Omega}, \tag{1.5}$$

whose stability can be easily determined. Here $\Omega = \delta W_{21} W_{12}$ and $\alpha = \delta^2 W_{21} W_{12} - 1$. The eigenvalues associated to the Jacobi matrix for this system for $P_0^* = (0, 0)$ are

$$\lambda_{\pm} = -1 \pm \delta \sqrt{W_{12} W_{21}}, \tag{1.6}$$

and thus this point will be stable if $\delta \sqrt{W_{12} W_{21}} < 1$. There is an exchange of stability and P_1^* becomes stable when the previous condition does not hold (i. e. a transcritical bifurcation takes place) [22].

When self-interactions are also considered (i. e. $W_{ii} \neq 0$) several attractors can be present as a consequence of the competition between positive feedbacks and mutual inhibition. One particular case is given by networks such that the matrix of connections $\mathbf{W}$ is symmetric, of the form:

$$\mathbf{W} = \begin{pmatrix} \gamma & \beta \\ \beta & \gamma \end{pmatrix}, \tag{1.7}$$

with $\beta \in \mathbb{R}$ and $\gamma > 0$. In other words, when there is self-activation by both genes plus cross interactions which can be positive or negative. The later is a very common situation in real morphogenetic processes and is strongly related with the process of competition between species in ecosystems.

The stability analysis of this general problem can be performed by using the general Jacobi matrix:

$$\mathbf{L} = \begin{pmatrix} \alpha\delta/\epsilon_{12}^2 - 1 & \alpha\beta/\epsilon_{12}^2 \\ \alpha\beta/\epsilon_{21}^2 & \alpha\delta/\epsilon_{21}^2 - 1 \end{pmatrix} \tag{1.8}$$

where $\epsilon_{ij} \equiv 1 + \alpha g_i^* + \beta g_j^*$. For $\beta > 0$, the mutual reinforcement between both genes leads to the same state (indicated as *homogeneous* in Figure 1.4). Here $g_1^* = g_2^* = [\delta(\alpha + \beta) - 1]/(\alpha + \beta)$ and it is stable (this point disappears at $\beta = (1 - \delta\gamma)/\delta$). For $\beta < -1$, the self-interaction is unable to sustain gene activity and it decays to zero. Finally, an interesting domain is observed for $(1 - \delta\gamma)/\delta > \beta > 0$, where three attractors are present (the previous one, where both coexist, and two exclusion points). In Figure 1.4(b) we show an example of the flow field for the 3-attractor domain. We can see that there are three basins of attraction associated to each possible final state (fixed point).

These results, in particular the presence of multiple attractors for some parameter ranges, are specially important within the context of development [20, 21]. In many cases the behavior of cells that become differentiated is very similar to that of a switch. By depending on initial conditions or external perturbations, which might emerge from some other genes in the networks, the system can reach one or another basin of attraction and thus a different final state. More importantly, it has been shown that some well-defined, small sets of interacting genes (so-called modules), are responsible for specific spatial patterns emerging in morphogenetic processes [20, 21]. As a consequence, not only single genes, but modules, can be the target of selection.

1.2.2 Random networks

Beyond the specific wiring diagrams that can be considered in small-sized genetic nets, the study of large-N nets has been dominated by randomly-wired systems [16]. Here genes are

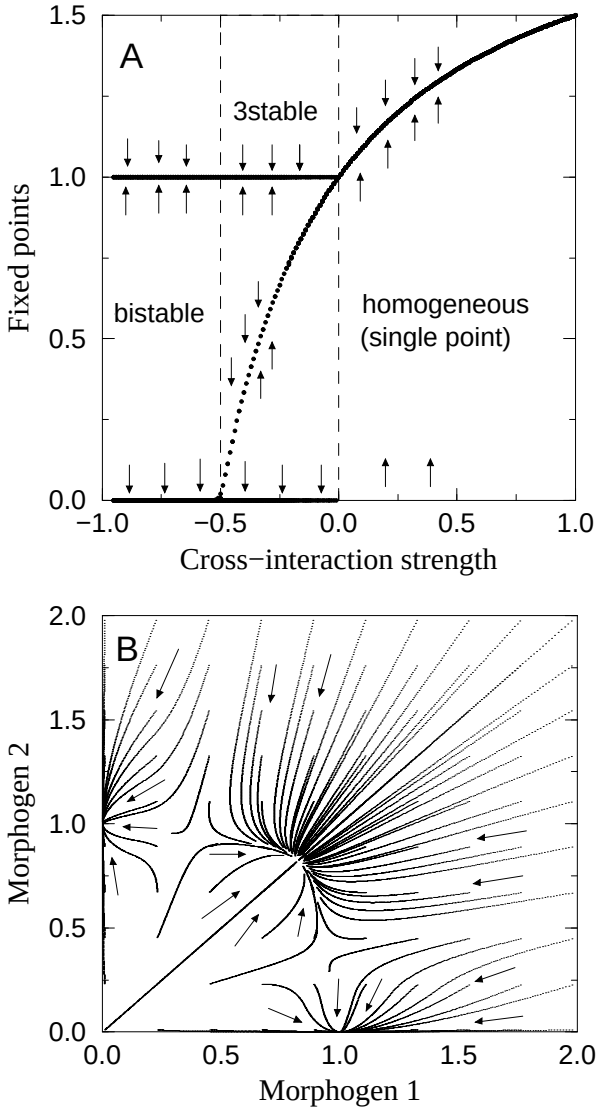

Figure 1.4: Multistability in gene network models: (a) bifurcation diagram for the two-gene network model with a symmetric matrix. Here $\delta = 2$ and $\gamma = 1$. Three basic domains are involved (see text); (b) flow diagram of the model for $\beta = -0.15$, in the three-attractor domain, indicated as 3stable in (a).

connected at random, with an average number of z connections per gene. An extensive literature on random Boolean networks has shown that a number of generic features are characteristic of these nets as a consequence of the presence of phase transition phenomena in random

graphs [23].

In order to illustrate this idea, let us consider a graph $\Omega_{n,p}$ that consists of n nodes joined by links with some probability p. Specifically, each possible link between two given nodes occurs with a probability p. The average number of links (also called the average *degree*) of a given node will be $z = np$, and it can be easily shown that the probability $P(k)$ that a node has a degree k (it is connected to k other nodes) follows a Poisson distribution,

$$P(k) = e^{-z}\frac{z^k}{k!}. \tag{1.9}$$

This so-called Erdös-Renyi (ER) random graph [24] will be fairly well characterized by an average degree

$$\langle k \rangle = \sum_k kP(k) = z, \tag{1.10}$$

where $P(k)$ shows a peak. The distribution $P(k)$ is in this sense a single-scaled distribution [25] and an example is shown in Figure 1.5(a).

The ER model displays a phase transition at a given critical average degree $z_c = 1$ [23, 26]. At this critical point, a *giant component* forms: for $z > z_c$ a large fraction of the nodes are connected in a single cluster, whereas for $z < z_c$ the system is fragmented into small subwebs. This type of random model has been used in different contexts, including ecological, genetic, metabolic, and neural networks [26]. The importance of this phase transition is obvious in terms of the collective properties that arise at the critical point: communication among the whole system becomes possible, and thus information can flow from the units to the whole system and back. Besides, the transition occurs suddenly and implies an innovation. No less important, it takes place at a low cost in terms of the number of required links ($\sim N$).

The ER model can be extended to directed graphs and has been analyzed by Kauffman within the context of genetic regulatory networks [16]. In the language presented in section 1.2, this will correspond to a network in which genes are randomly connected, and regulated by an average of z other genes. This means that the $N \times N$ matrix $\mathbf{W} = \{W_{ij}\}$ will have zN^2 nonzero elements, distributed at random. The probability that a gene is regulated by exactly k other genes will be then given by the distribution (1.9). Beyond the specific time-dependent features associated to the particular model chosen, one important characteristic of these systems is the presence of the percolation threshold: once a critical average connectivity $z_c = 1$ (the ratio of directed links to genes) is reached, the system becomes suddenly connected. Below the critical threshold the system is essentially disconnected and thus changes in a given gene cannot propagate to the rest of the system. The presence of the percolation threshold allows the system to exhibit a complex dynamical behavior, including deterministic chaos, Figure 1.5(b).

One consequence of these models (but strongly tied to the topological properties of sparse random graphs) is that a high diversity of attractors compatible with a high degree of homeostasis seems to naturally emerge close to the percolation threshold. However, early evidence indicated that the degree distributions that characterize *real* genetic nets are far from Poissonian. Actually, as we will see in section 1.4, the topology of real networks strongly departs from the Erdös-Renyi scenario.

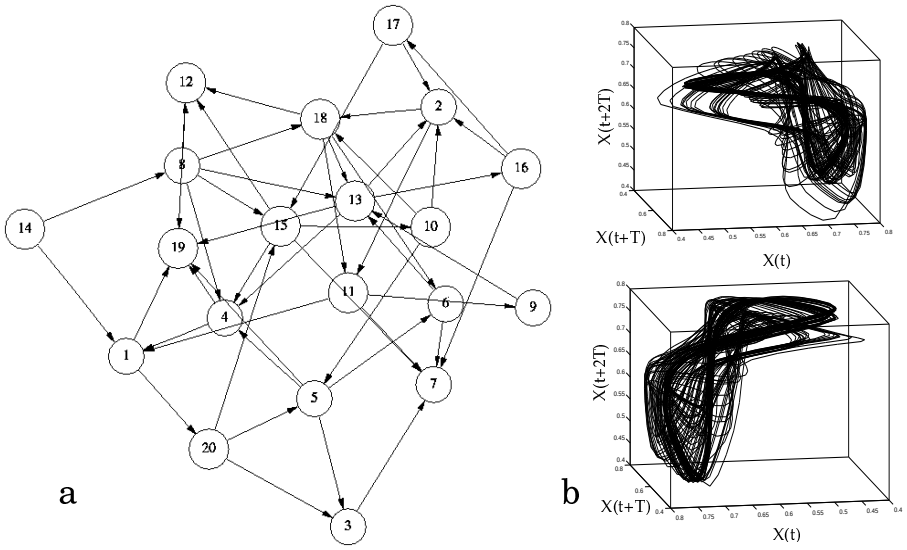

Figure 1.5: (a) An example of a directed random network with Poissonian structure. Here each node is a gene in a model gene network and arrows indicate the regulatory connections. This type of graph is characterized by an average degree z; together with the appropriate nonlinear coupling among genes, it can generate different types of dynamical patterns, including deterministic chaos. An example of the strange attractors obtained from these nets is shown in (b) in two different views.

1.3 Three interconnected levels of cellular nets

Gene regulation takes place at different levels and involves the participation of proteins. The whole cellular network includes three levels of integration:

- The genome, and the regulation pathways defined by interactions among genes;
- The proteome, defined by the set of proteins and their interactions; and
- The metabolic network, also under the control of proteins that operate as enzymes.

Unlike the relatively unchanging genome, the dynamic proteome changes through time in response to intra- and extracellular environmental signals. The proteome is particularly important. Proteins unify genome structure on the one hand and functional biology on the other: they are both the products of genes and regulate reactions or pathways.

Complicating the study of gene function is the fact that multiple proteins can arise from a single gene. In eukaryotic organisms, genes appear fragmented into pieces (exons) separated by non-coding domains (introns). After transcription, the resulting messenger RNA (mRNA) is generated by the excision and elimination of introns followed by the joining of exons. This process is called *splicing*. Once the mRNA is formed, it will be translated into a protein by the translation machinery.

A very important feature is that splicing can occur in different ways so that different sets of exons are joined together. In this way, different mRNA's (and thus different proteins) are produced. The combinatorial potential of this so-called *alternative splicing* is obvious. In some cases, thousands of different proteins are potentially available for a given gene.

Alternative splicing expands genome complexity in an extraordinary fashion. In this context, although the genomes of complex organisms might not strongly differ in terms of their number of genes, the underlying proteome complexity can be very different. As will be discussed in the next sections, the actual structure of protein networks is shown to be strongly heterogeneous and shares several previously unsuspected traits with many different systems.

1.4 Small world graphs and scale-free nets

The analysis of the topological structure of protein interaction maps (in the budding yeast *Saccharomyces cerevisiae* and other simple organisms) revealed a surprising result: the protein-protein interaction net shares some universal features with the topological organization of other complex nets, both natural and artificial, ranging form technological networks [27, 25, 28, 29], neural networks [30], metabolic pathways [31, 32, 33], and food webs [34, 35] to the human language graph [36]. These studies actually offer the first global view of the proteome map and show that it strongly departs from the simple Erdös-Renyi scenario.

The first feature characteristic of the proteome map is that the probability $P(k)$ that a given protein interacts with other k proteins has a *scale-free* (SF) nature, i.e. it follows a power law, $P(k) \sim k^{-\gamma}$, with a sharp exponential cut-off for large k. Thus most proteins have a small number of links with other proteins and a few of them are highly connected (*hubs*). Those last ones are likely to be very important to cell function [37, 38, 6].

The second feature is the presence of the so-called *small world* (SW) property [30, 39, 40]. Small world graphs have a number of surprising features that make them specially relevant to understand how interactions among individuals, metabolites, or species lead to the robustness and homeostasis observed in nature. The SW pattern can be detected from the analysis of two basic statistical properties of the network[1]: (a) the *clustering coefficient* C and (b) the *average path length* $\bar{\ell}$.

The proteome graph (see Figure 1.6) is defined by a pair $\Omega_p = (W_p, E_p)$, where $W_p = \{p_i\}, (i = 1, ..., N)$ is the set of N proteins (nodes) and $E_p = \{\{p_i, p_j\}\}$ is the set of edges/connections between proteins. The *adjacency matrix* ξ_{ij} indicates that an interaction exists between proteins $p_i, p_j \in W_p$ ($\xi_{ij} = 1$) or that the interaction is absent ($\xi_{ij} = 0$). Two connected proteins are thus called *adjacent* and the *degree* k_i of a given protein is the number of edges that connect it with other proteins. Let us consider the adjacency matrix and indicate by $\Gamma_i = \{p_j \,|\, \xi_{ij} = 1\}$ the set of nearest neighbors of a protein $p_i \in W_p$. The clustering coefficient for this protein is defined as the ratio between the actual number of connections between the proteins $p_j \in \Gamma_i$, and the total possible number of connections, $k_i(k_i - 1)/2$ [30]

[1] Since the proteome map is a disconnected network, these quantities are actually defined on the *giant component*, defined as the largest cluster of connected nodes in the network [23].

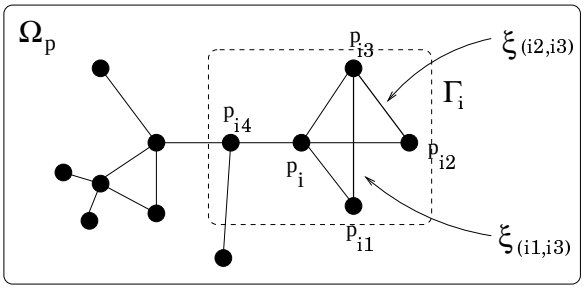

Figure 1.6: Measuring the clustering from a proteome graph Ω_p. Here each node (black circles) is a protein and physical interactions are indicated by means of edges connecting nodes.

(see Figure. 1.6). Denoting

$$\mathcal{L}_i = \sum_{j=1}^{N} \xi_{ij} \left[\sum_{k \in \Gamma_i} \xi_{jk} \right], \tag{1.11}$$

we define the clustering coefficient of the i-th protein as

$$C(i) = \frac{2\mathcal{L}_i}{k_i(k_i - 1)}, \tag{1.12}$$

where k_i is the degree of the i-th protein. The clustering coefficient is defined as the average of $C(i)$ over all the proteins,

$$C = \frac{1}{N} \sum_{i=1}^{N} C(i). \tag{1.13}$$

The average path length $\bar{\ell}$ is defined as follows: Given two proteins $p_i, p_j \in W_p$, let ℓ_{ij} be the length of the shortest path connecting these two proteins, following the links present in the network. The average path length $\bar{\ell}$ is defined as:

$$\bar{\ell} = \frac{2}{N(N-1)} \sum_{i<j}^{N} \ell_{ij}. \tag{1.14}$$

For the ER graph, we have a clustering coefficient inversely proportional to the network size, $C_{\rm ER} \approx z/N$; this is a very small quantity, that tends to zero for large networks. The average path length, on the other hand, is proportional to the logarithm of the network size $\bar{\ell}_{ER} \approx \log(N)/\log(z)$. At the other extreme, regular lattices with only nearest-neighbor connections among units exhibit a long average path length. Graphs with SW structure are characterized by a high clustering, $C \gg C_{\rm ER}$, while possessing an average path comparable to an ER graph with the same average connectivity and number of nodes, $\bar{\ell} \approx \bar{\ell}_{\rm ER}$.

The experimental observations on the proteome map can be summarized as follows:

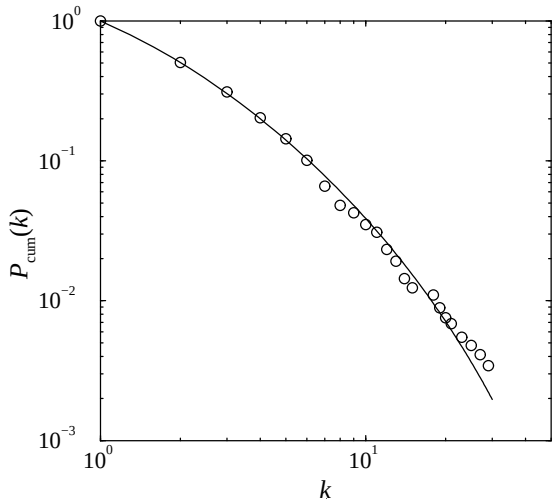

Figure 1.7: (b) Cumulated degree distribution for the yeast proteome map from Ref. [37]. The degree distribution has been fitted to the scaling behavior $P(k) \approx (k_0 + k)^{-\gamma} e^{-k/k_c}$, with an exponent $\gamma \simeq 2.6$ and a sharp cut-off $k_c \simeq 15$.

1. The proteome map is a sparse graph, with a small average number of links per protein. In Ref. [41] an average connectivity $z \sim 1.9 - 2.3$ was reported for the proteome map of *S. cerevisiae*. This observation is also consistent with the study of the global organization of the *E. coli* gene network from available information on transcriptional regulation [42].

2. It exhibits a SW pattern, different from the properties displayed by purely random (ER) graphs. In particular, Ref. [41] reported the values $C = 2.2 \times 10^{-2}$ and $\bar{\ell} = 7.14$, to be compared with the values corresponding to an ER network with comparable size and average connectivity, $C_{\rm ER} = 1 \times 10^{-3}$ and $\bar{\ell}_{ER} = 8.0$.

3. The degree distribution of links follows a power-law with a well-defined cut-off. To be more precise, Jeong *et al.* [37] reported a functional form for the degree distribution of *S. cerevisiae*

$$P(k) \simeq (k_0 + k)^{-\gamma} e^{-k/k_c}. \tag{1.15}$$

Parameters reported in Ref. [37] are $k_0 \simeq 1$, $\gamma \approx 2.4$ and a cut-off $k_c \approx 20$. In Figure 1.7 we check this functional dependence on the cumulated degree distribution of the protein map[2] used in Ref. [37]. A fit to the form (1.15) yields the values $k_0 \simeq 1.1$, $k_c \simeq 15$, and $\gamma = 2.6 \pm 0.2$, compatible with the results found in [37, 41]. This particular form of the degree distribution could have adaptive significance as a source of robustness against mutations.

The highly heterogeneous character of these maps has important consequences within the context of molecular cell biology [32, 6]. It indicates that the evolution of proteome/genome

[2]Data available at the web site `http://www.nd.edu/~networks/database/index.html`.

complexity has been driven towards a well-defined topological pattern that provides the substrate for an extraordinary homeostatic stability against random mutational events.

1.5 Scale-free proteomes: gene duplication models

Several models have been proposed in order to explain the regularities displayed by the proteome map [44, 45, 46]. These models of proteome evolution are based on a gene duplication plus rewiring process that includes the basic ingredients of proteome growth and intends to reproduce the previous set of observations. The first component of the models allows the system to grow by means of the copy process of previous units (together with their wiring). The second introduces novelty by means of changes in the wiring pattern, usually constrained to the newly created genes. This constraint is required if we assume that conservation of gene (protein) interactions is due to functional restrictions and that further changes in the regulation map are limited. Such constraint would be strongly relaxed when involving a newly created (and redundant) unit. The models proposed so far are intended to capture the *topological* properties of the proteome map. No explicit functionality is included in the description of the proteins and this is certainly a drawback. But by ignoring the specific features of the protein-protein interactions and the underlying regulation dynamics, one can explore the question of how much the network topology is due to the duplication and diversification processes.

In this chapter we will focus in particular in the model described in Refs. [45, 46]. This model considers single-gene duplications, which occur in most cases due to unequal crossover [47], plus re-wiring. Multiple duplications should be considered in future extensions of these models: molecular evidence shows that even whole-genome duplications have actually occurred in *S. cerevisiae* [48] (see also Ref. [49]). Re-wiring has also been used in dynamical models of the evolution of robustness in complex organisms [50].

The proteome graph at any given step t (i.e. after t duplications) will be indicated as $\Omega_p(t)$. The rules of the model, summarized in Figure 1.8, are implemented as follows. Each time step: (a) one node in the graph is randomly chosen and duplicated; (b) the links emerging from the new generated node are removed with probability δ; (c) finally, new links (not previously present) can be created between the new node and all the rest of the nodes with probability α. Step (a) implements gene duplication, in which both the original and the replicated proteins retain the same structural properties and, consequently, the same set of interactions. The rewiring steps (b) and (c) implement the possible mutations of the replicated gene, which translate into the deletion and addition of interactions, with different probabilities.

1.5.1 Mean-field rate equation for the average connectivity

Since the model just presented has two free parameters, namely the deletion probability δ and the addition probability α, one preliminary task is to constrain their possible values by using the available empirical data. One average property that can be determined is the evolution of the average number of interactions per protein/gene through time, which can be compared with the evidence from real proteomes [37, 41], as well as recent analysis of large-scale perturbation experiments [51].

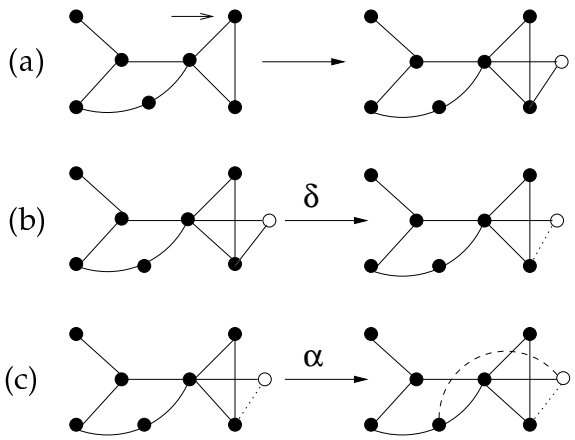

Figure 1.8: Growing network by duplication of nodes. First (a) duplication occurs after randomly selecting a node (arrow). The links from the newly created node (white) now can experience deletion (b) and new links can be created (c); these events occur with probabilities δ and α, respectively.

Let us indicate by z_N and L_N the average connectivity of the system and its number links, respectively, when it is composed by N proteins. These magnitudes satisfy the relation $L_N = z_N N/2$. It is easy to check (see also Ref. [44]) that, at a mean-field level, that number of links L_N fulfill the following rate equation

$$L_{N+1} = L_N + z_N + \alpha(N - z_N) - \delta z_N, \tag{1.16}$$

where the last two terms correspond to the addition of links to a fraction α to the $N - z_N$ units not connected to the duplicated node, plus the deletion of any of the new z_N links, with probability δ. Using the continuous approximation

$$\frac{dz_N}{dN} \simeq z_{N+1} - z_N, \tag{1.17}$$

Eq. (1.16) can be written

$$\frac{dz_N}{dN} = \frac{1}{N}\left[z_N + 2\alpha(N - z_N) - 2\delta z_N\right], \tag{1.18}$$

whose solution is

$$z_N = \frac{\alpha}{\alpha + \delta} N + \left(z_1 - \frac{\alpha}{\alpha + \delta}\right) N^{\Gamma}, \tag{1.19}$$

where $\Gamma = 1 - 2(\alpha + \delta)$ and z_1 is the initial connectivity at $N = 1$. For any constant value of α and δ this model leads to an increasing connectivity through time. In order to have a finite z in the limit of large N, one possible solution is to impose an addition rate α that is a function of the size of the network, with the form

$$\alpha(N) = \frac{\beta}{N}, \tag{1.20}$$

where β is a constant. That is, the rate of addition of new links (the establishment of new viable interactions between proteins) is inversely proportional to the network size, and thus much smaller than the deletion rate δ, in agreement with the rates observed in [41]. In this case, for large N, the differential rate equation (1.18) equation takes the form

$$\frac{dz_N}{dN} = \frac{1}{N}(1-2\delta)z_N + \frac{2\beta}{N}. \tag{1.21}$$

The solution of this equation is

$$z_N = \frac{2\beta}{2\delta - 1} + \left(z_1 - \frac{2\beta}{2\delta - 1}\right) N^{1-2\delta}. \tag{1.22}$$

For $\delta > 1/2$ a finite connectivity is reached in the limit of a large network,

$$z \equiv \lim_{N\to\infty} z_N = \frac{2\beta}{2\delta - 1}. \tag{1.23}$$

In order to reduce the number of independent parameters of the model, Ref. [45] used the available experimental data to estimate the average degree z and the ratio of addition and deletion rates in the yeast proteome, α/δ [41] to find a relation between β and δ, which, together with Eq. (1.23), yields a numerical estimate of β and δ. Since it is clear that this estimate is strongly dependent on the assumed value α/δ, Ref. [46] followed a more pragmatical approach, considering a δ-dependent model and fixing the actual value of δ by comparing numerical simulations with experimental data.

1.5.2 Rate equation for the node distribution n_k

The rate equation approach to evolving networks [52] can be fruitfully applied to the proteome model under consideration [46]. This approach focuses on the time evolution of the number $n_k(t)$ of nodes in the network with exactly k links at time t. Defining our network by means of the set of numbers $n_k(t)$, we have that the total number of nodes N is given by

$$N = \sum_k n_k, \tag{1.24}$$

while the total number of links is given by

$$L = \frac{1}{2}\sum_k k n_k. \tag{1.25}$$

Time is divided into periods. In each period, $t \to t+1$, one node is duplicated at random, so that $N \to N+1$. If, after each duplication, there is a probability δ to delete each link from the just-duplicated node, the probability of increasing the number of nodes at degree k, by direct duplication without link deletion, is given by

$$\Pr_{\text{self,dup}}[n_k \to n_k + 1] = \frac{n_k}{N}(1 - k\delta). \tag{1.26}$$

On the other hand, a node of degree k can be created from the duplication of a node of degree $k+1$ in which a link is deleted, contributing with a probability

$$\Pr_{\text{above,dup}}[n_k \to n_k + 1] = \frac{n_{k+1}}{N}(k+1)\delta. \tag{1.27}$$

The probability of degree change, from duplication of a node connected to a degree-k node, is given by:

$$\Pr_{\text{other,dup}}[(n_{k-1}, n_k) \to (n_{k-1} - 1, n_k + 1)] = \frac{n_{k-1}}{N}(k-1)(1-\delta). \tag{1.28}$$

Finally, in the same period, we proceed to add $N - k_d$ links with probability $\alpha = \beta/N$, where k_d is the connectivity of the just duplicated node. In the limit $N \gg k_d$, we can simply consider the addition of $N\alpha = \beta$ new links to the graph. When this last step is performed with the *correlated* prescription given for the model (i.e. adding links from the duplicated node to the rest of the nodes in the graph), it leads to a nonlocal rate equation for the functions n_k [46]. For the sake of simplicity, we will consider now the simpler case of a *uncorrelated* addition of links (new links created between any two nodes in the graph). However, it can be proved that both prescriptions lead qualitatively to similar results [46].

The case of uncorrelated addition of links can be represented as the distribution of $2\alpha N$ new link ends among the N nodes in the network. This event contributes with a probability

$$\Pr_{\text{add}}[(n_k, n_{k+1}) \to (n_k - 1, n_{k+1} + 1)] = \frac{n_k}{N} 2\alpha N = \frac{n_k}{N} 2\beta, \tag{1.29}$$

The probabilities (1.26), (1.27), (1.28), and (1.29) define the rate equation for the connectivity distribution

$$\begin{aligned} \frac{dn_k(t)}{dt} &= \frac{n_k}{N} + \frac{\delta}{N}[(k+1)n_{k+1} - kn_k] + \frac{1-\delta}{N}[(k-1)n_{k-1} - kn_k] \\ &+ \frac{2\beta}{N}[n_{k-1} - n_k]. \end{aligned} \tag{1.30}$$

Since each time step a new node is added, Eq. (1.30) satisfies the condition

$$\frac{dN}{dt} = \sum_k \frac{dn_k(t)}{dt} = 1, \tag{1.31}$$

that yields the expected result $N(t) = N_0 + t$, where N_0 is the initial number of nodes in the network. In order to solve Eq. (1.30), we impose the homogeneous condition on the population number

$$n_k(t) = N(t)p_k \simeq tp_k, \tag{1.32}$$

where p_k is the probability of finding a node of connectivity k, which we assume to be independent of time. With this approximation, the rate equation reads

$$(k+1)\delta p_{k+1} - (k + 2\beta)p_k + [(k-1)(1-\delta) + 2\beta]p_{k-1} = 0. \tag{1.33}$$

Eq. (1.33) can be solved using the generating functional method [53]. Let us define the the generating functional

$$\phi(x) = \sum_k x^k p_k. \tag{1.34}$$

Introducing this definition into Eq. (1.33), we obtain an equation for $\phi(x)$, whose solution is

$$\phi(x) = \left(\frac{\delta - x(1-\delta)}{2\delta - 1}\right)^{-2\beta/(1-\delta)}. \tag{1.35}$$

Knowing $\phi(x)$ we can compute immediately the average connectivity

$$z = \sum_k k p_k \equiv x\frac{d\phi(x)}{dx}\bigg|_{x=1} = \frac{2\beta}{2\delta - 1}, \tag{1.36}$$

in agreement with the mean-field prediction of Eq. (1.23). On the other hand, performing a Taylor expansion of $\phi(x)$ around $x = 0$ we can obtain p_k as

$$p_k = \frac{1}{k!}\frac{d\phi(x)}{dx}\bigg|_{x=0}. \tag{1.37}$$

Applying this formula to the function (1.35), and using Stirling's approximation for large k, we can obtain the asymptotic behavior of p_k, given by:

$$p_k \sim (k_0 + k)^{-\gamma} e^{-k/k_c}, \tag{1.38}$$

with

$$\gamma = -k_0 = 1 - \frac{2\beta}{1-\delta}, \qquad k_c = \frac{1}{\ln\left(\frac{\delta}{\delta - 1}\right)}. \tag{1.39}$$

As we can observe from the previous result, we recover the same functional form experimentally observed in [37]. However, it is important to notice that for all the parameter range in which the exponential cut-off k_c is well-defined, we obtain a value of the degree exponent, as given by Eq. (1.39), that is $\gamma \leq 1$. The same result holds when considering the rate equation for the correlated model, in which the link addition is fully correlated with the new duplicated node [46]. This result is unsatisfactory, because it does not correspond with the results from numerical simulations of the model [46]. This discrepancy is explained by the fact that the $N \to \infty$ solution presented has only meaning for $\delta > 1/2$ (see Eq. (1.36)). Yet the master equation was defined on the basis of an independent-event approximation that only makes sense for $\delta \ll 1$. The master equation itself should become valid for $\delta \to 0$, but then the convergence results assumed at $N \to \infty$ seem questionable, as indicated by the fact that we get an analytic, but negative, z.

There is, however, something qualitative still to be learned from these equations, in the neighborhood of $\delta \sim 1/2$, small β. This is a neighborhood where the convergence results at large N still give sensible answers, even if they are not quantitatively correct due to marginal approximations in the underlying master equation. Yet at the same time, since this is the smallest value of δ where we can get answers, it is the one where the master equation we have constructed is likely to be the best approximation to the much more complicated true equation (one with frequent coupled events).

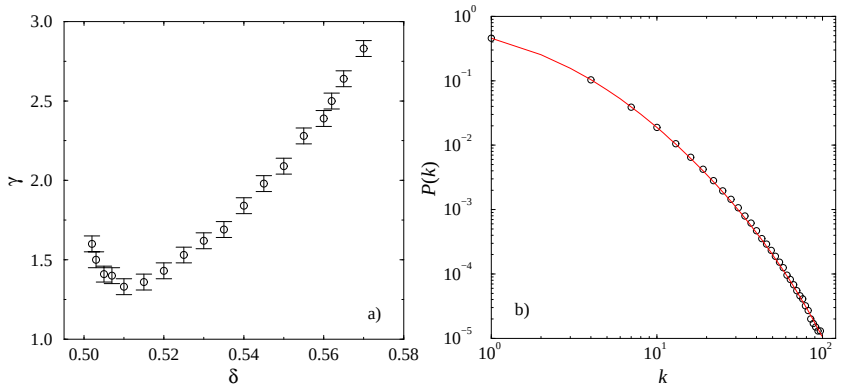

Figure 1.9: a) Degree exponent γ as a function of the deletion rate δ from computer simulations of the proteome model with average connectivity $z = 2.5$. Network size $N = 2 \times 10^3$. The degree distributions is averaged over 1000 different network realizations. b) Degree distribution for the same model, $\delta = 0.562$, averaged over 10000 different network realizations. The distribution can be fitted to the form $P(k) \approx (k_0 + k)^{-\gamma} e^{-k/k_c}$, with an exponent $\gamma = 2.5 \pm 0.1$ and a sharp cut-off $k_c \simeq 37$.

1.5.3 Numerical simulations

The proteome model defined in section 1.5 depends effectively on two independent parameters: the average connectivity of the network z and the deletion rate of newly created links δ, being the addition rate β computed from Eq. (1.23). The average connectivity can be estimated from the experimental results from real proteome maps. The data analyzed in Ref. [37] yields a value $z \simeq 2.40$. As a safe estimate, one can impose the value $z = 2.5$ [46], and consider values $\delta > 1/2$, in accordance with Eq. (1.23). In spite of the drawbacks of the analytical study of the model, section 1.5.2, one should expect the model to yield, for each value of δ, the functional form Eq. (1.38) for the degree distribution, with a degree exponent γ which is a function of δ (for a fixed average connectivity $z = 2.5$). From numerical simulations of the model one can then compute the function $\gamma(\delta)$ and select the value of δ that yields a degree exponent in agreement with the experimental observations. Fig 1.9(a) shows values of γ estimated from the functional form (1.38) for the degree distribution obtained from computer simulations of model, averaging over 1000 network of size $N = 2 \times 10^3$ nodes, of the same order of those found in the maps analyzed in Ref. [37]. Apart form a concave region for δ very close to $1/2$, γ is an increasing function of δ. The value of δ yielding the degree exponent closest to the experimentally observed one is then

$$\delta = 0.562. \tag{1.40}$$

In Figure 1.10(a) we show the topology of the giant component of a typical realization of the network model of size $N = 2 \times 10^3$. This Figure clearly resembles the giant component of a real yeast networks, as we can see comparing with Figure 1.10(b)[3]; we can appreciate the

[3]Figure kindly provided by W. Basalaj (see `http://www.cl.cam.uk/~wb204/GD99/#Mewes`).

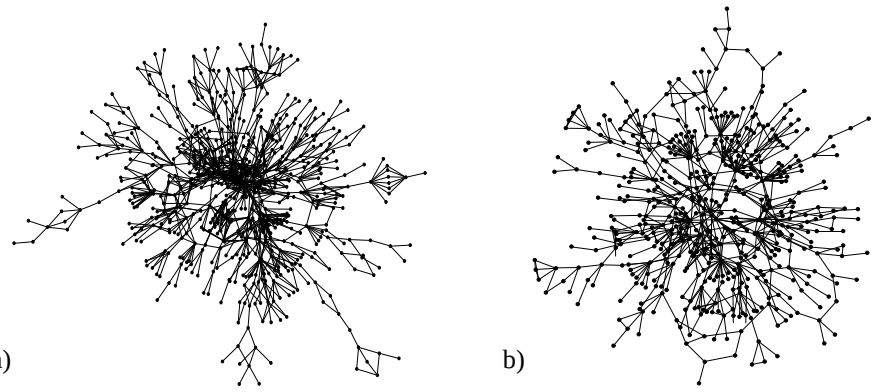

Figure 1.10: a) Topology of the giant component of the map obtained with the proteome model with parameters $\langle k \rangle = 2.5$ and $\delta = 0.565$. Network size $N = 2 \times 10^3$. b) Topology of a real yeast proteome map obtained from the MIPS database [43].

presence of a few highly connected hubs plus many nodes with a relatively small number of connections, in close resemblance of the real proteome map. On the other hand, Figure 1.9(b) shows the connectivity $P(k)$ obtained for networks of size $N = 2 \times 10^3$, averaged of 10000 realizations, for $\delta = 0.562$. In this Figure we observe that the resulting connectivity distribution can be fitted to a power-law with an exponential cut-off, of the form given by Eq. (1.38), with parameters $\gamma = 2.5 \pm 0.1$ and $k_c \simeq 37$, in fair agreement with the measurements reported by [41] and [37].

Finally, Table 1.1 reports the values of z, γ, C, and $\bar{\ell}$ obtained for the proteome model, compared with the values reported for the yeast *S. cerevisiae* by Ref. [41], those calculated for the map used in Ref. [37], and those corresponding to an ER random graph with size and average connectivity comparable with both the model and the real data. All the magnitudes displayed by the model compare quite well with the values measured for the yeast, and represent a further confirmation of the SW conjecture for the protein networks advanced by [41].

Table 1.1: Comparison between the observed regularities in the yeast proteome reported by Wagner [41], those calculated from the proteome map used by Jeong *et al.* [37], the model predictions with $N = 2000$, $\delta = 0.562$ and $z = 2.50$, and a ER network with the same size and connectivity as the model.

	Wagner's data	Jeong's data	Proteome model	ER model
z	1.83	2.40	2.4 ± 0.6	2.50 ± 0.05
γ	2.5	2.4	2.5 ± 0.1	—
C	2.2×10^{-2}	7.1×10^{-2}	1.0×10^{-2}	1×10^{-3}
$\bar{\ell}$	7.14	6.81	5.5 ± 0.7	8.0 ± 0.2

1.6 Discussion

Simple models of complex biological interactions have been used through the last decades as powerful metaphors of natural complexity. Networks pervade biology and there is little doubt that the untangling biological complexity demands a considerable degree of simplification. This view works well when generic mechanisms are at work: percolation close to criticiality in random graphs would be a perfect example in this context. Since information transfer (network communication) is a key property to all biosystems, reaching a threshold in connectivity allows information to propagate in a very effective way under a low wiring cost.

Similar principles might be operating in technology graphs [28, 54] and the striking similarities between man-made networks (such as electronic circuits or software graphs) and natural webs suggests that an organizing principle involving optimal communication might be at work in both types of systems. This seems a reasonable possibility, since the cost of wiring is an important constraint in both cases. For technology graphs, however, random failure tipically leads to collapse and thus there is no robustness associated to the scale-free architecture. Biological systems might have taken advantage of the SF patterns that arise from optimization of path length under low cost [55] and make use of the source of robustness *for free* that might be generated.

As it occurs with many other aspects of biological complexity, historic constraints play an important role in shaping network topology. Not surprisingly, some of the oldest actors in the metabolic scene seem to be highly connected, thus suggesting a leading role of preferential attachment [26] at least at early stages of the evolution of metabolism. But the proteome map is a very large web incorporating a large amount of plasticity that might have been tuned through evolution in order to reach optimaly wired pathways. Future research will provide a new perspective on how biological nets get organized through evolution and what are the contributions of emergence, selection, and tinkering to network biocomplexity.

Acknowledgements

We thank the members of the Complex Systems Lab for useful discussions. This work has been partially supported by the European Network Contract No. ERBFMRXCT980183, the European Commission - Fet Open project COSIN IST-2001-33555, a grant PB97-0693 and by the Santa Fe Institute (R. V. S.). R.P.-S. acknowledges financial support from the Ministerio de Ciencia y Tecnología (Spain).

References

[1] D. Bray. Protein molecules as computational elements in living cells. *Nature* **376**, 307–312 (1995).

[2] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).

[3] P. Ross-Macdonald, P. S. R. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K. H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, M. Heldtman, F. K. Nelson,

H. Iwasaki, K. Hager, M. Gerstein, P. Miller, G. S. Roeder, and M. Snyder. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).

[4] A. Wagner. Robustness against mutations in genetic networks of yeast. *Nature Genet.* **24**, 355–361 (2000).

[5] P. H. Kussie, S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine and N. P. Pavletich. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **274**, 948–953 (1996) .

[6] B. Vogelstein, D. Lane, and A. J. Levine. Surfing the p53 network. *Nature* **408**, 307–310 (2000).

[7] S. Jin. Identification and characterization of a p53 homologue in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **97**, 7301–7306 (2000).

[8] K. W. Kohn Molecular interaction map of the mammalian cell cylce control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703-2734 (1999).

[9] R. J. Williams ans N. D. Martinez. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).

[10] H. Lodish, A. Berk, S. L. Zipursky, and P. Matsudaira, *Molecular Cell Biology*, (W. H. Freeman, New York, 2000). 4th edition.

[11] W. J. Gehring, *Master Control Genes in Development and Evolution*, (Yale University Press, New Haven, 1998).

[12] J, Hasty, D. McMillen, F. Isaacs and J. J. Collins Computational studies of gene regulatory networks: in numero molecular biology. *Nature Reviews Genet.* **2**, 268–279 (2001).

[13] P. Smolen, D. A. Baxter and J. H. Byrne Mathematical modeling of gene networks. *Neuron* **26**, 567–580 (2000).

[14] B. Goodwin, *Temporal organization in cells*, (Academic Press, New York, 1963).

[15] J. E. Lewis and L. Glass Steady states, limit cycles and chaos in models of complex biological networks. *Int. J. Bif. Chaos* **1**, 477–483 (1991).

[16] S. A. Kauffman, *Origins of Order*, (Oxford University Press, New York, 1993).

[17] S. B. Carroll. Chance and necessity: the evolution of morphological complexity and diversity. *Nature* **409**, 1102–1109 (2000).

[18] M. Laurent and N. Kellershohn Multistability: a major means of differentiation and evolution in biological systems. *Trends Biochem. Sci.* **24**, 418–422 (1999).

[19] M. Ptashne, *A Genetic Switch* (Blackwell, Cambridge, 1992).

[20] I. Salazar, J. Garcia-Fernandez and R. V. Solé Gene networks capable of pattern formation: from induction to reaction-diffusion. *J. Theor. Biol.* **205**, 587–603 (2000).

[21] R. V. Solé, I. Salazar and J. Garcia-Fernandez Common Pattern Formation, Modularity and Phase Transitions in a Gene Network Model of Morphogenesis. *Physica A* **305**, 640-647 (2002).

[22] J. M. T. Thompson and H. B. Stewart, *Nonlinear dynamics and chaos*, (John Wiley & Sons, New York, 1986).

[23] B. Bollobás, *Random Graphs*, (Academic Press, London, 1985).

[24] P. Erdös and P. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).

[25] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97**, 11149–11152 (2000).

[26] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

[27] R. A. Albert, H. Jeong, and A.-L. Barabási. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).

[28] R. Ferrer i Cancho, C. Janssen, and R. V. Solé. The topology of technology graphs: small world pattern in electronic circuits. *Phys. Rev. E* **63**, 32767 (2001).

[29] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.* **87**, 258701 (2001).

[30] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).

[31] D. Fell and A. Wagner. The small world of metabolism. *Nature Biotech.* **18**, 1121 (2000).

[32] H. Jeong, B. Tombor, R. Albert, Z. N.Oltvai, and A.-L. Barabasi. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2001).

[33] J. Podani, Z. Oltvai, H. Jeong, B. Tombor, A.-L. Barabási, and E. Szathmáry. Comparable system-level organization of Archaea and Eukaryotes. *Nature Genetics* **29**, 54–56 (2001).

[34] J. M. Montoya and R. V. Solé. Small world patters in fodd webs. *J. Theor. Biol.* **214**, 405–412 (2002).

[35] R. J. Williams, N. D. Martinez, E. L. Berlow, J. A. Dunne, and A.-L. Barabási. Two degrees of separation in complex food webs, (2001). Santa Fe working paper 01-07-036.

[36] R. Ferrer i Cancho, C. Janssen, and R. V. Solé. The small world of human language. *Procs. Roy. Soc. London B* **268**, 2261–2266 (2001).

[37] H. Jeong, S. Mason, A. L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature* **411**, 41 (2001).

[38] S. Maslov and K. Sneppen Specificity and stability in topology of protein networks. *Science* **296**, 910-913 (2002).

[39] D. J. Watts, *Small Worlds*, (Princeton University Press, Princeton, 1999).

[40] M. E. J. Newman Models of the Small World. *J. Stat. Phys.* **101**, 819–841 (2000).

[41] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18**, 1283–1292 (2001).

[42] D. Thieffry, A. M. Huerta, E. Pérez-Rueda, and J. Collado-Vives. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *escherichia coli*. *BioEssays* **20**, 433–440 (1998).

[43] H. W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. Mips: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**,44–48 (1999).

[44] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani. Modelling of protein interaction networks, (2001). cond-mat/0108043.

[45] R. V. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler. A model of large-scale proteome evolution. *Adv. Complex. Syst.* **5**, 43–54 (2002).

[46] R. Pastor-Satorras, E. Smith, and R. V. Solé. Evolving protein interaction networks through gene duplication, (2002). Santa Fe working paper 02-02-008.

[47] S. Ohno, *Evolution by gene duplication*, (Springer, Berlin, 1970).

[48] K. H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).

[49] A. Wagner. Evolution of gene networks by gene duplications: A mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA* **91**, 4387–4391 (1994).

[50] S. Bornholdt and K. Sneppen. Robustness as an evolutionary principle. *Proc. Roy. Soc. Lond. B* **267**, 2281–2286 (2000).

[51] A. Wagner. Estimating coarse gene network structure from large-scale gene perturbation data, (2001). Santa Fe working paper 01-09-051.

[52] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629 (2000).

[53] C. W. Gardiner, *Handbook of stochastic methods*, (Springer, Berlin, 1985). 2nd edition.

[54] S. Valverde, R. Ferrer Cancho and R. V. Solé. Scale-free networks from optimal design. Santa Fe working paper 02-04-019.

[55] R. Ferrer Cancho and R. V. Solé. Optmization in Complex Networks. Santa Fe working paper 01-11-068.