Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 222 (2003) 199–210
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

Evolving protein interaction networks through gene duplication

Romualdo Pastor-Satorras[a], Eric Smith[b], Ricard V. Solé[c,*]

[a] Dept. de Fisica, FEN, Universitat Politècnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain
[b] Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA
[c] ICREA-Complex Systems Lab, Universitat Pampeu Fabra, Dr Aiguader 80, 08003 Barcelona, Spain

Received 8 March 2002; received in revised form 12 December 2002; accepted 20 December 2002

Abstract

The topology of the proteome map revealed by recent large-scale hybridization methods has shown that the distribution of protein–protein interactions is highly heterogeneous, with many proteins having few edges while a few of them are heavily connected. This particular topology is shared by other cellular networks, such as metabolic pathways, and it has been suggested to be responsible for the high mutational homeostasis displayed by the genome of some organisms. In this paper we explore a recent model of proteome evolution that has been shown to reproduce many of the features displayed by its real counterparts. The model is based on gene duplication plus re-wiring of the newly created genes. The statistical features displayed by the proteome of well-known organisms are reproduced and suggest that the overall topology of the protein maps naturally emerges from the two leading mechanisms considered by the model.

Keywords: Genome; Proteome; Evolution; Complex networks; Sealing

1. Introduction

Since the discovery of the structure of the DNA molecule, a dominant view of molecular biology has been the understanding of the microscopic mechanisms operating at the gene level. Some authors have indeed defined molecular cell biology as an explanation of organisms and cells in terms of their individual molecules (Lodish et al., 2000). This so-called *reductionist* view has been extremely successful and has widely enlarged our view of genetics and evolution at the smallest scales. In approaching the richness of biocomplexity in this way, we might, however, ignore the other side of the coin: the presence of higher-order phenomena beyond the molecular level. This other view takes into account the interactions among components as an essential part of the whole picture and suggests that there exist *emergent properties*, not reducible to the properties displayed by the individual components (Goodwin, 1994). The debate between both schools goes back to the early origins of molecular biology (Monod, 1970).

Our perspective of molecular biology might be changing and the emerging picture might help to reach a more balanced interaction between both views. Two important findings help to see how collective properties might play a leading role. The first is the observation of the extraordinary resilience exhibited by some simple organisms against gene removal. Experiments with systematic mutagenesis in yeast *Saccharomyces cerevisiae* have shown a great tolerance of this organism to gene removal (Ross-Macdonald et al., 1999; Wagner, 2000). These and other studies carried out in order to explore the minimum limits allowed to genome size suggest that many genes might not play a key phenotypic role, being somehow functionally replaced by other genes. Secondly, the network perspective of gene and protein systems is becoming more and more accepted as new data accumulate. In particular, it is becoming obvious that not only genes, but also interactions among specific groups of genes (modules) have been conserved through evolution (Hartwell et al., 1999). In this context, subsets of complex gene networks are also the target of selective forces.

*Corresponding author. ICREA-Complex Systems Lab, Universitat Pampeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain. Tel.: +34-93-5422821; fax: +34-93-2213237.
E-mail address: ricard.sole@cexs.upf.es (R.V. Solé).

0022-5193/03/$ - see front matter
doi:10.1016/S0022-5193(03)00028-6

Recent large-scale studies of the global properties of the yeast proteome reinforce the relevance of the network perspective (Jeong et al., 2001a; Wagner, 2001; Gavin, 2002; Ho, 2002). These studies have revealed that the available data from protein–protein interaction networks in the yeast *S. cerevisiae* share some unexpected features with other complex networks (Jeong et al., 2001a; Wagner, 2001; Goh et al., 2002). In particular, these are very heterogeneous networks, whose degree distribution $P(k)$ (i.e. the probability that a protein interacts with any other k proteins) displays a scale-free behavior, $P(k) \approx k^{-\gamma}$, with a characteristic exponent $\gamma \approx 2.5$, for a certain range of values of k, and with a well-defined cut-off for large k. Additionally, they also display the so-called small-world (SW) effect: they are highly clustered (each vertex has a well-defined neighborhood of "close" vertices) but the minimum distance between any two randomly chosen vertices in the graph is short, a characteristic feature of random graphs (Watts and Strogatz, 1998; Watts, 1999). Scale-free (SF) networks appear to be present in many natural and artificial systems (Bornholdt and Schuster, 2002), ranging from technological networks (Albert et al., 2000; Amaral et al., 2000; Ferrer i Cancho et al., 2001a; Pastor-Satorras et al., 2001), neural networks (Watts and Strogatz, 1998; Stephan et al., 2000), metabolic pathways (Fell and Wagner, 2000; Jeong et al., 2001b; Podani et al., 2001), and food webs (Montoya and Solé, 2002; Williams et al., 2001) to the human language graph (Ferrer i Cancho et al., 2001b). It is remarkable, in particular, that the exponents observed in Internet, metabolic, and protein networks are very similar. This fact hints towards the presence of common principles of organization, a finding which might have deep consequences in our understanding of how large-scale nets emerge through evolution.

Previous studies on protein networks have emphasized dynamical or computational aspects of interacting proteins as well as their potential links with other classes of nets, such as neural nets (Bray, 1995). The importance of allosteric interactions (and their non-linear character) was early highlighted as an essential piece in the understanding of cell biology and as a step towards a general systems theory of biocomplexity (Monod, 1970). Here, however, we are mainly interested in the topological properties derived from the process of proteome evolution. These properties can be summarized as follows (Jeong et al., 2001a; Wagner, 2001): (1) the proteome map is a sparse graph indicating a small average number of edges per protein. This observation is also consistent with the study of the global organization of the *Escherichia coli* gene network from available information on transcriptional regulation (Thieffry et al., 1998); (2) it exhibits a small-world pattern, very different from the properties displayed by purely random (Poissonian) graphs (Bollobás, 1985); and (3) the degree distribution is a power law with a well-defined cut-off.

In this paper we present a model of proteome evolution based on a gene duplication plus re-wiring process that includes the basic ingredients of proteome growth and intends to reproduce the previous set of observations. The first component of the model 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, constrained in our approach to the newly created genes (Solé et al., 2002b). 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 evolution of the globin gene family provides an example of genome evolution by gene duplication. Here the primitive, single-chain globin molecule (found in many insects and primitive fish) is closely related to the hemoglobin molecule present in higher vertebrates. The hemoglobin is composed of a tetramer formed by two copies of two slightly different globin chains. About 500 million years ago, a series of gene duplications and mutations took place, establishing the two different globin genes which allow, through the interaction of their protein products, a more efficient transport of oxygen.

The model does not include functionality or dynamics in the proteins involved. It is a topological-based approximation to the overall features of the proteome graph which aims to capture some of the (possibly) generic features of real proteome evolution. A preliminary account of the present model was previously given in a short communication (Solé et al., 2002b). It is also worth noting the work by Vázquez et al. (2003), in which a related model of proteome evolution, showing multifractal degree properties, is described and analysed.

This paper is organized as follows. In Section 2 we review the known topological properties of proteome networks, obtained by several authors by analysing the published proteome maps of the yeast *S. cerevisiae*. In Section 3 we describe our model of proteome growth. The main ingredients of the model are protein duplication plus correlated random re-wiring. Sections 4 and 5 are devoted to an analytical study of the model. In Section 4 we discuss a mean-field approximation for the evolution of the average degree, that will allow us to restrict the range of values of the model's parameters, while Section 5 presents a study of the rate equation for the vertex distribution n_k within an approximation that imposes an uncorrelated re-wiring of connections after each vertex duplication. Our study is completed in Section 6 by means of computer simulations. In Section 7 we present a discussion of our results.

2. Topological properties of real proteome maps

Protein–protein interaction maps have been studied, at different levels, in a variety of organisms including viruses (Bartel et al., 1996; Flajolet et al., 2000; McCraith et al., 2000), prokaryotes (Rain et al., 2001), yeast (Ito et al., 2000), and multicellular organisms such as *Caenorhabitis elegans* (Walhout et al., 2000). Previous studies have mainly used the so-called two-hybrid assay (Fromont-Racine et al., 1997), based on the properties of site-specific transcriptional activators. Although differences exist between different two-hybrid projects (Hazbun and Fields, 2001), the statistical patterns used in our study seem to be robust. Recent systematic analyses of protein complexes by means of mass spectrometry provide very similar results, together with a better understanding of the internal organization of protein complexes (Kumar and Snyder, 2001).

From a statistical point of view, protein–protein interaction maps can be viewed as a random network (Bollobás, 1985), in which the vertices represent the proteins and an edge between two vertices indicates the presence of an interaction between the respective proteins. Mathematically, the proteome graph is defined by a pair $\Omega_p = (W_p, E_p)$, where $W_p = \{p_i\}$, $(i = 1, \ldots, N)$ is the set of N proteins 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 $p_i, p_j \in \Omega_p$ ($\xi_{ij} = 1$) or that the interaction is absent ($\xi_{ij} = 0$). Two connected proteins are thus called *adjacent* and the *degree* of a given protein is the number of edges that connect it with other proteins. The underlying key unit here is the so-called protein *domain*. A domain is defined as a subpart of a protein chain that can fold independently into a stable structure. These domains act as the modules from which complex proteins are built, and different domains are associated to different functions. A given protein can have domains involved in regulatory activity and others in catalytic activity. Domain networks have been recently explored, revealing the presence of scaling at different levels (Wuchty, 2001; Rzhetsky and Gomez, 2001): here two domains are considered to be *connected* if they are simultaneously present in the same protein. In this context, it is remarkable that scaling laws seem to be universal, from bacteria to metazoa, suggesting common scenarios of proteome domain evolution. In our model (see below) no explicit domains are included, but they are implicitly introduced in terms of edges among proteins (Fig. 1).

The network representation of the protein interactions, shown in Fig. 2(a),[1] reveals a very complex topology, characterized by the presence of several highly

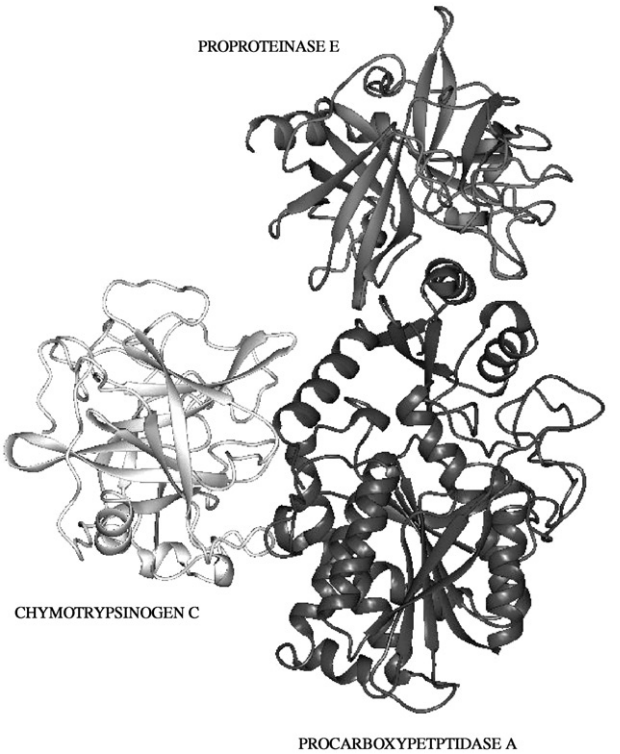

Fig. 1. An example of three interacting proteins describing a ternary complex. Here interactions involve physical contact between residues of different molecules. (Figure courtesy of Baldo Oliva.)

connected hubs, while most of the proteins have very few connections. The network topology can be statistically characterized by means of the degree distribution $P(k)$, defined as the probability that any vertex is connected to exactly k other vertices. The analysis of the protein map from the yeast *S. cerevisiae*, containing 1870 vertices and 2240 edges, corresponding to an average degree (average number of edges emanating from a vertex) $\langle k \rangle = 2.40$, shows that the degree distribution can be fitted to a power law with an exponential cut-off, of the form:

$$P(k) \sim (k_0 + k)^{-\gamma} e^{-k/k_c}. \quad (1)$$

The estimated values for the yeast are $k_0 \simeq 1$, $\gamma \simeq 2.4$ and $k_c \simeq 20$ (Jeong et al., 2001a). That is, the protein map is a SF network with a characteristic cut-off. This value is confirmed by the independent analysis of Wagner (2001), who found a power-law behavior with $\gamma \simeq 2.5$ for a relatively smaller protein map (985 vertices with average degree $\langle k \rangle = 1.83$). In Fig. 2(b) we have checked this functional dependence on the cumulated degree distribution of the protein map used in Jeong et al. (2001a) (available at the web site http://www.nd.edu/~networks/database/index.html). A fit to form (1) 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 Jeong et al. (2001a) and Wagner (2001).

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

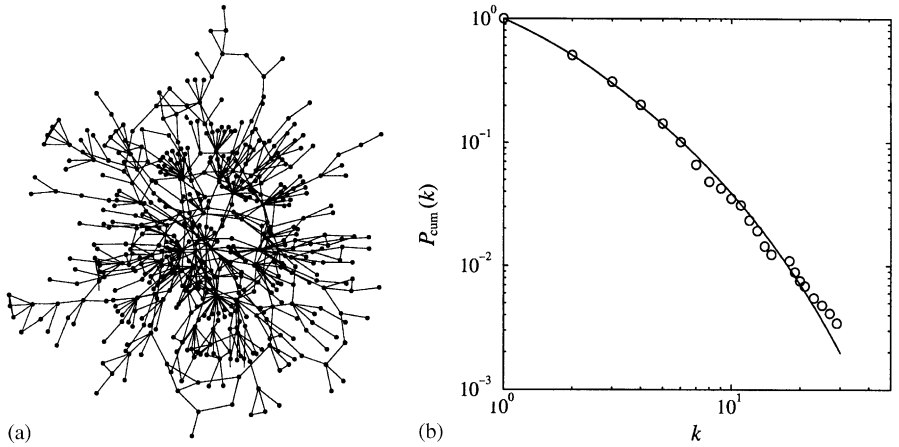

Fig. 2. (a) Topology of a real yeast proteome map obtained from the MIPS database (Mewes et al., 1999). (b) Cumulative degree distribution for the yeast proteome map from Jeong et al. (2001a). The proteome map is available at the web site http://www.nd.edu/~networks/database/index.html. The degree distribution 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 $k_c \simeq 15$.

An additional observation from Wagner's study of the yeast proteome is the presence of SW properties (Watts and Strogatz, 1998). The SW pattern can be detected from the analysis of two basic statistical quantities: the *clustering coefficient* C and the *average path length* $\bar{\ell}$. Since the proteome map is a disconnected network, these quantities are defined on the *giant component* Ω_∞, defined as the largest cluster of connected vertices in the network (Bollobás, 1985). Let us consider the adjacency matrix of the giant component, ξ_{ij}^∞ and indicate by $\Gamma_i = \{p_j \,|\, \xi_{ij}^\infty = 1\}$ the set of nearest neighbors of a protein $p_i \in \Omega_\infty$. The clustering coefficient for this protein is defined as the number of connections between the proteins $p_j \in \Gamma_i$ (Watts and Strogatz, 1998). Denoting

$$\mathscr{L}_i = \sum_{j=1}^{N_\infty} \xi_{ij}^\infty \left[\sum_{k \in \Gamma_i} \xi_{jk}^\infty \right], \tag{2}$$

where N_∞ is the size of the giant component, we define the clustering coefficient of the ith protein as

$$C(i) = \frac{2\mathscr{L}_i}{k_i(k_i - 1)}, \tag{3}$$

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

$$C = \frac{1}{N_\infty} \sum_{i=1}^{N_\infty} C(i), \tag{4}$$

and it provides a measure of the average fraction of pairs of neighbors of a vertex that are also neighbors of each other.

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

$$\bar{\ell} = \frac{2}{N_\infty(N_\infty - 1)} \sum_{i<j}^{N_\infty} \ell_{ij}. \tag{5}$$

Random graphs, where vertices are randomly connected with a given probability p (Bollobás, 1985), have a clustering coefficient inversely proportional to the network size, $C^{rand} \approx \langle k \rangle / N$, and an average path length proportional to the logarithm of the network size, $\bar{\ell}^{rand} \approx \log N / \log \langle k \rangle$. At the other extreme, regular lattices with only nearest-neighbor connections among units are typically clustered and exhibit a large average path length. Graphs with SW structure are characterized by a high clustering, $C \gg C^{rand}$, while possessing an average path comparable with a random graph with the same average degree and number of vertices, $\bar{\ell} \approx \bar{\ell}^{\,rand}$.

In Table 1 we summarize the most relevant results for the proteome map of the yeast, as reported in Wagner (2001). In order to compare with other results, we report the values we have calculated for the map used in Jeong et al. (2001a), as well as for a random graph with size and average degree comparable with the real data. These values support the conjecture of the SW properties of the protein network put forward in Wagner (2001).

3. Proteome growth model

In this work we will consider the scenario of single-gene duplications. Although multiple-gene duplications should also be taken into account (even whole genome duplication), here we restrict our attention to the most common ones (Ohno, 1970), which are known to occur due to unequal crossover (Patthy, 1999). After duplication of a single, randomly chosen gene, new connections

Table 1
Comparison between the observed regularities in the yeast proteome reported by Wagner (2001), those calculated from the proteome map used in Jeong et al. (2001a), the model predictions with $N = 2000$, $\delta = 0.562$ and average degree $\langle k \rangle = 2.5$ (see Section 6), and a random network with the same size and average degree as the model

	Wagner (2001)	Map from Jeong et al. (2001a)	Network model	Random network
$\langle k \rangle$	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

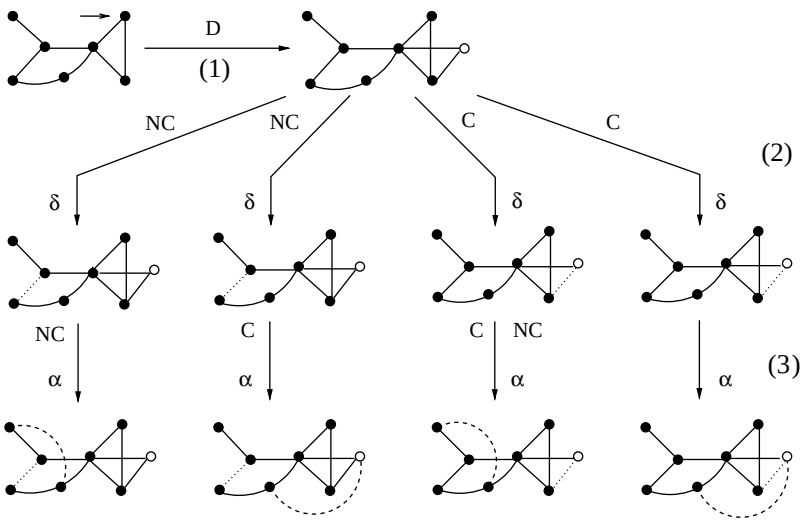

Fig. 3. Rules of proteome growth in the four possible scenarios. First, (1) duplication occurs after randomly selecting a vertex (small arrow). Then (2) deletion of connections occurs with probability δ. This event can be correlated (C) when the deleted edges are connected to the newly generated vertex or uncorrelated (NC), when all edges are considered for deletion. Finally (3) new connections are generated with probability α, again in a correlated or uncorrelated way.

can be added and previous connections deleted. Both rewiring rules can be implemented in a *correlated* or *uncorrelated* manner. The first involves changes that affect the just duplicated gene and its connections. The second involves any edge in the network. Both processes (creation and deletion of edges) might be associated or not to the newly created unit. Four possible combinations are thus allowed in principle: (1) correlated creation and deletion of edges; (2) correlated creation and uncorrelated deletion of edges; (3) uncorrelated creation and correlated deletion of edges; (4) uncorrelated creation and deletion of edges. The rules associated with each variation of the model are summarized in Fig. 3.

In this work we will focus in the first variation of the model, in which created and deleted edges occur in relation with the newly duplicated vertex. The reason to consider correlations has to do with the assumption that the evolutionary significance of gene duplication lies in the fact that changes in the newly created genes can lead to the emergence of novelty (Patthy, 1999). After gene duplication, one of the two copies becomes redundant and either one of them becomes non-functional (i.e. a pseudogene) or accumulates molecular changes that provide a new function. The new function might be very different. An example is provided by mouse lysozyme genes. One of them has a digestive function in the intestine and the second has a bactericide action in myeolid tissues. Strong divergences from the original function displayed by the ancestor can develop. Moreover, from a numerical point of view, the analysis of the models in which creation or deletion of edges is uncorrelated yield results which are in disagreement with experimental observations. Additionally, we do not consider differences associated to the number of edges of a given chosen vertex. In this sense, evidence from proteome data indicate that the degree of well-conserved proteins is negatively correlated with their rate of evolution (Fraser et al., 2002).

The model we will consider is defined by the following rules. We start from a set N_0 of connected vertices, and each time step we perform the following operations:

(i) One vertex of the graph is selected at random and duplicated.
(ii) The edges emanating from the newly generated vertex are removed with probability δ. Available data indicate that δ is very large.

(iii) New edges (not previously present after the duplication step) are created between the new vertex and all the rest of the vertices with probability α. Available data indicate that this number is very small (much smaller than δ).

Step (i) 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 re-wiring steps (ii) and (iii) implement the possible mutations of the replicated gene, which translate into the deletion and addition of interactions with different proteins, respectively.

The model we have just defined is 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, we can explore the question of how much the network topology is due to the duplication and diversification processes.

4. Analytical study of the model: mean-field rate equation for the average degree

Since we have two free parameters in our model, namely the deletion probability δ and the addition probability α, we should first constrain their possible values by using the available empirical data. One first 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 (Jeong et al., 2001a; Wagner, 2001), as well as recent analysis of large-scale perturbation experiments. This can be done for any model with vertex duplication plus addition/deletion of vertices by considering the discrete dynamics of the number or edges L_N at a given step N, where N is the number of vertices in the network (see also Vázquez et al. (2003)). In general, we can write the evolution equation

$$L_{N+1} = L_N + K_N + \phi_a(K_N, L_N) - \phi_d(K_N, L_N), \tag{6}$$

where $K_N = 2L_N/N$ indicates average degree at the Nth duplication event, and ϕ_a and ϕ_d stand for the general rates of addition/deletion of vertices, respectively. Here different functional forms might be chosen, including rates of change that depend on the degree of the vertex, as suggested by some studies (Fraser et al., 2002). Although duplication rate would in principle depend on the number of edges too, available data indicate that no such association seems to be present (Wagner, 2001).

For the particular case of the model defined in the previous section, the rate equation takes the form:

$$L_{N+1} = L_N + K_N + \alpha(N - K_N) - \delta K_N, \tag{7}$$

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

$$\frac{dK_N}{dN} \simeq K_{N+1} - K_N. \tag{8}$$

Eq. (7) can be written

$$\frac{dK_N}{dN} = \frac{1}{N}[K_N + 2\alpha(N - K_N) - 2\delta K_N] \tag{9}$$

whose solution is

$$K_N = \frac{\alpha}{\alpha+\delta}N + \left(K_1 - \frac{\alpha}{\alpha+\delta}\right)N^{\Gamma}, \tag{10}$$

where $\Gamma = 1 - 2(\alpha + \delta)$ and K_1 is the initial degree at $N = 1$. For any value of α and δ this version leads to an increasing degree through time. In this context, the $\alpha = \delta = 0$ case would correspond to a pure duplication process, where K_N increases linearly with time.

In order to have a final sparse graph with low numbers of edges per protein, we need to consider two possible scenarios. The first would assume fixed α and δ values and a finite N, that we take as the proteome size. Assuming that $\delta + \alpha > 1/2$ in order to ensure $\Gamma < 0$, the asymptotic behavior of K_N is dominated by the first, linear term. If the desired degree is indicated as K^*, the required number of vertices N^* will be

$$N^* = \left\lceil \frac{\alpha+\delta}{\alpha} K^* \right\rceil, \tag{11}$$

where $\lceil x \rceil$ indicates the integer part of x.

Another possibility is to assume that the rate of edge creation scales as the inverse of N, i.e. $\alpha = \beta/N$, where $\beta > 0$ is some constant. This would be interpreted in terms of underlying constraints to the number of edges imposed by some network-level property (such as efficient communication at low cost; see Ferrer i Cancho and Solé, 2001). In any case, addition of new edges occurs at a very slow rate and under our assumptions a fixed, very small number of edges will be introduced at each step. Our analysis, as well as other independent studies (Vázquez et al., 2003; Kim et al., 2002), indicate that the key statistical features of the final proteome map (such as the scaling exponent γ) do not depend on the rate of edge addition.

Here the rate of addition of new edges (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 experimentally observed. Using this scaling form, the rate equation for K_N reads as

$$\frac{dK_N}{dN} = \frac{1}{N}(1 - 2\delta)K_N + \frac{2\beta}{N} - \frac{2\beta K_N}{N^2}. \tag{12}$$

The time-dependent solution of this equation is

$$K_N = N^{1-2\delta} e^{2\beta/N}[C + (2\beta)^{2\delta}\Gamma(1-2\delta, 2\beta/N)], \quad (13)$$

where C is an integration constant and $\Gamma(a,z)$ is the incomplete Gamma function (Abramowitz and Stegun, 1972). For large values of N (small z) we can use the Taylor expansion of $\Gamma(a,z)$, given by

$$\Gamma(a,z) = \Gamma(a) - z^a \sum_{m=0}^{\infty} \frac{(-z)^m}{(m+a)m!} \quad (14)$$

that yields

$$K_N = N^{1-2\delta} e^{2\beta/N}[C + (2\beta)^{2\delta}\Gamma(1-2\delta)] - 2\beta e^{2\beta/N} \sum_{m=0}^{\infty} \frac{(-2\beta/N)^m}{(m+1-2\delta)m!}. \quad (15)$$

For $\delta > 1/2$, a finite average degree is reached at infinite N,

$$K_\infty = \lim_{N\to\infty} K_N = \frac{2\beta}{2\delta - 1}. \quad (16)$$

This is thus consistent with the data analysis by Wagner (2001). Eq. (16) can be used to restrict the number of independent parameters of the model, by fixing K_∞ to the values experimentally found in real proteome maps. Thus, we can fix the value of β by

$$\beta = (\delta - 1/2)K_\infty \equiv (\delta - 1/2)\langle k \rangle. \quad (17)$$

5. Analytical study of the model: rate equation for the vertex distribution n_k

The rate equation approach to evolving networks (Krapivsky et al., 2000) can be fruitfully applied to the proteome model under consideration. This approach focuses on the time evolution of the number $n_k(t)$ of vertices in the network with exactly k edges at time t. Defining our network by the set of numbers $n_k(t)$, we have that the total number of vertices N is given by

$$N = \sum_k n_k, \quad (18)$$

while the total number of edges is given by

$$L = \frac{1}{2}\sum_k k n_k, \quad (19)$$

since the sum over vertex connections double-counts edges.

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

$$Pr_{self,\,dup}[n_k \to n_k + 1] = \frac{n_k}{N}(1 - k\delta). \quad (20)$$

In this expression n_k/N represents the probability of selecting a vertex of degree k and $1 - k\delta$ is the probability of preserving all edges in the just duplicated vertex. It is important to note that in Eq. (20) we are ignoring the possibility of deleting more than one edge in each duplication event, which will contribute with an amount proportional to δ^2 or smaller. Obviously, this approximation is correct for small δ. We will see that this fact has important consequences when interpreting the results obtained in this section.

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

$$Pr_{above,\,dup}[n_k \to n_k + 1] = \frac{n_{k+1}}{N}(k+1)\delta. \quad (21)$$

In this expression, the factor $(k+1)\delta$ represents the probability of deleting one of the $k+1$ connections of the duplicated vertex. The probability of degree change, from duplication of a vertex connected to a degree-k vertex, is given by

$$Pr_{other,\,dup}[(n_{k-1}, n_k) \to (n_{k-1} - 1, n_k + 1)] = \frac{n_{k-1}}{N}(k-1)(1-\delta), \quad (22)$$

because kn_k is the total number of vertices connected to all vertices of degree k. In Eq. (22) we have corrected for the probability δ that the crucial connecting edge was deleted.

Finally, in the same period, we proceed to add $N - k_d$ edges with probability $\alpha = \beta/N$, where k_d is the degree of the just duplicated vertex. In the limit $N \gg k_d$, we can simply consider the addition of $N\alpha$ new edges to the graph. When this last step is performed with the correlated rule (i.e. adding edges from the duplicated vertex to the rest of the edges in the graph), it leads to a non-local rate equation for the functions n_k. For the sake of simplicity, we will consider now the simpler case of a uncorrelated addition of edges (new edges created between any two vertices in the graph).

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

$$Pr_{add}[(n_k, n_{k+1}) \to (n_k - 1, n_{k+1} + 1)] = \frac{n_k}{N} 2\alpha N = \frac{n_k}{N} 2\beta. \quad (23)$$

Probabilities (20)–(23) define the rate equation for the degree distribution:

$$\frac{dn_k(t)}{dt} = \frac{n_k}{N} + \frac{\delta}{N}[(k+1)n_{k+1} - kn_k] + \frac{1-\delta}{N}[(k-1)n_{k-1} - kn_k] + \frac{2\beta}{N}[n_{k-1} - n_k]. \quad (24)$$

The point to note in Eq. (24) is the first term proportional to n_k/N. This is the unaltered duplication event, which can create a vertex of degree n_k only by duplicating another such vertex. It is separated from the edge addition probabilities in the rest of the terms, because for re-wired edges, there is no correlation between the likelihood that a vertex of degree k will be created by duplication, and that it will be gained or lost by edge addition. Since in each time step a new vertex is added, Eq. (24) satisfies the condition

$$\frac{dN}{dt} = \sum_k \frac{dn_k(t)}{dt} = 1 \tag{25}$$

that yields the expected result $N(t) = N_0 + t$, where N_0 is the initial number of vertices in the network. In order to solve Eq. (24), we impose the homogenous condition on the population number

$$n_k(t) = N(t)p_k \simeq tp_k, \tag{26}$$

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

$$(k+1)\delta p_{k+1} - (k+2\beta)p_k + [(k-1)(1-\delta) + 2\beta]p_{k-1} = 0. \tag{27}$$

Eq. (27) can be solved using the generating functional method (Gardiner, 1985). Let us define the generating functional

$$\phi(x) = \sum_k x^k p_k. \tag{28}$$

In terms of ϕ, Eq. (27) can be written as

$$[(1-\delta)x^2 - x + \delta]\frac{d\phi(x)}{dx} + 2\beta(x-1)\phi(x) = 0. \tag{29}$$

The solution of this last equation, with the boundary condition $\phi(1) = \sum_k p_k = 1$, is

$$\phi(x) = \left(\frac{\delta - x(1-\delta)}{2\delta - 1}\right)^{-2\beta/(1-\delta)}. \tag{30}$$

Knowing the form of $\phi(x)$ we can compute immediately the average degree

$$\langle k \rangle = \sum_k k p_k \equiv x\frac{d\phi(x)}{dx}\bigg|_{x=1} = \frac{2\beta}{2\delta - 1}, \tag{31}$$

in agreement with the mean-field prediction of Eq. (16).

On the other hand, performing a Taylor expansion of $\phi(x)$ around $x = 0$ we can obtain p_k as

$$p_k = \frac{1}{k!}\phi^{(k)}(0), \tag{32}$$

where $\phi^{(k)}(x)$ is the kth derivative of $\phi(x)$. Applying this formula on function (30), we are led to

$$p_k = \left(\frac{2\delta - 1}{\delta}\right)^{2\beta/(1-\delta)} \times \frac{1}{\Gamma(\frac{2\beta}{1-\delta})}\frac{\Gamma(\frac{2\beta}{1-\delta} + k)}{k!}\left(\frac{\delta}{1-\delta}\right)^{-k}. \tag{33}$$

By using Stirling's approximation, we can obtain the asymptotic behavior of p_k for large k, which is given by

$$p_k \sim (k_0 + k)^{-\gamma} e^{-k/k_c} \tag{34}$$

with

$$\gamma = -k_0 = 1 - \frac{2\beta}{1-\delta}, \quad k_c = \frac{1}{\ln(\frac{\delta}{\delta - 1})}. \tag{35}$$

As we observe from Eq. (34), we recover the same functional form experimentally observed in Jeong et al. (2001a). However, it is important to note 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. (35), that is $\gamma \leqslant 1$. This result is unsatisfactory, because, as we will see in the next section, it does not correspond with the results from numerical simulations of the model. This discrepancy is explained by the fact that the $N \to \infty$ solution that we have constructed only applies for $\delta > 1/2$ (see Eq. (31)). Yet the master equation was defined on the basis of an independent-event approximation that only makes sense for $\delta \ll 1/2$.

In summary, the master equation correctly reproduces the scale-free distribution but does not give the appropriate exponent (something that the model correctly reproduces when simulated, see next section). The singular character of this model has been stressed in a recent study by Kim et al. (2002). In this sense, it has been shown that the fluctuations displayed by the model are very important. By following a somewhat different approach, these authors also obtained scale-free distributions with a characteristic exponent in the limit of large k (infinite large network)

$$\gamma(\delta) = 1 + \frac{1}{1-\delta} - (1-\delta)^{\gamma - 2}. \tag{36}$$

This relation yields an exponent γ surprisingly independent of the addition probability factor β. The solution of this equation for $\delta \approx 0.56$ is $\gamma \approx 2.7$, consistent with numerical simulations (see next section).

6. Numerical results

The proteome model defined in Section 3 depends effectively on two independent parameters: the average degree of the network $\langle k \rangle$ and the deletion rate of newly created edges δ; given these two parameters, the rate β can be computed from Eq. (17). The average

degree can be estimated from the experimental results from real proteome maps. Examination of Table 1 yields a value $\langle k \rangle \simeq 2.40$ from the data analysed in Jeong et al. (2001a). As a safe estimate, we impose the value $\langle k \rangle = 2.5$ in our model. In Solé et al. (2002b) the rate δ was roughly estimated from the experimentally calculated ratio of addition and deletion rates in the yeast proteome, α/δ (Wagner, 2001). However, it is clear that this estimate is strongly dependent on the assumed value α/δ. In this work we will consider instead the more general case of a δ-dependent model. Guided by the analytical study in Section 5, we should expect the model to yield, for each value of δ, the functional form Eq. (1) of the degree distribution, with a degree exponent γ which is a function of δ (for a fixed average degree $\langle k \rangle = 2.5$). From numerical simulations of the model we will compute the function $\gamma(\delta)$ and select the value of δ that yields a degree exponent in agreement with the experimental observations.

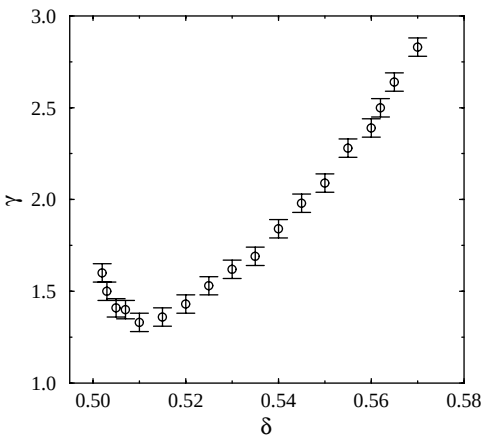

Fig. 4. Degree exponent γ as a function of the deletion rate δ from computer simulations of the proteome model with average degree $\langle k \rangle = 2.5$. Network size $N = 2 \times 10^3$. The degree distributions is averaged over 1000 different network realizations.

Simulations of the model start from a connected ring of $N_0 = 5$ vertices and proceed by iterating the rules of the model until the desired network size is achieved. Given the size of the maps analysed by Jeong et al. (2001a), we consider networks with $N = 2 \times 10^3$ vertices. In Fig. 4 we plot the values of γ estimated from functional form (1) for the degree distribution obtained from computer simulations of our model, averaging over 1000 network realizations. The exponent γ is computed performing a nonlinear regression of the corresponding degree distribution in the range $k \in [1, 80]$. In this figure we observe that, apart from a concave region for δ very close to $1/2$, γ is an increasing function of δ. We thus conclude that the value of δ yielding the degree exponent closest to the experimentally observed one is

$$\delta = 0.562. \tag{37}$$

We will use this value thorough the rest of the paper.

In Fig. 5(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 real yeast networks, as we can see comparing with Fig. 2(a); we can appreciate the presence of a few highly connected hubs plus many vertices with a relatively small number of connections. On the other hand, in Fig. 5(b) we plot the degree $P(k)$ obtained for networks of size $N = 2 \times 10^3$, averaged of 10 000 realizations. In this figure we observe that the resulting degree distribution can be fitted to a power law with an exponential cut-off, of the form given by Eq. (1), with parameters $\gamma = 2.5 \pm 0.1$ and $k_c \simeq 37$, in fair agreement with the measurements reported by Wagner (2001) and Jeong et al. (2001a) (see also Table 1). This value is also to be compared with the theoretical prediction from Kim et al. (2002), $\gamma \approx 2.7$. The slight

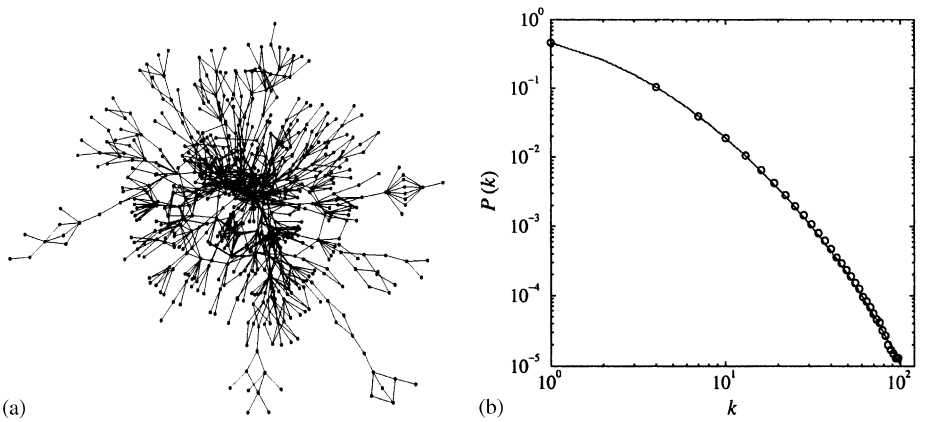

Fig. 5. (a) Topology of the giant component of the map obtained with the proteome model with parameters $\langle k \rangle = 2.5$ and $\delta = 0.562$. Network size $N = 2 \times 10^3$. (b) Degree distribution for the same model, averaged over 10 000 different network realizations.

discrepancy in the value is to be attributed to the small size of the simulated networks, comparable to the size of real proteome maps, while Eq. (36) is expected to be valid in the limit $N \to \infty$.

We have also computed the SW properties of the model. In Table 1 we report the values of $\langle k \rangle$, γ, C, and $\bar{\ell}$ obtained for our model, compared with the values reported for the yeast *S. cerevisiae* (Jeong et al., 2001a), and the values corresponding to a random graph with size and degree 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 Wagner (2001).

7. Discussion

In this paper a detailed analysis of a model of proteome evolution (Solé et al., 2002b) has been presented. The model is a simple approximation to the evolution of the real proteome map, and no functionality is considered (i.e. no cellular network dynamics is explicitly introduced). This simplification imposes some limitations to the conclusions reachable by our study. Nevertheless, the success in reproducing the observed statistical features of real interaction maps suggests that our mechanism is able to capture the essential ingredients that shape large-scale proteome evolution, at least those that can be extracted from topological data, without introducing selective forces that would fine-tune the proteome structure. In this context, it is important to mention that, regardless of the limitations and biases imposed by different large-scale molecular methods (from two-hybrid assays to mass spectrometry) there seems to be a strong consistency in the overall pattern that results from these different sources (Kumar and Snyder, 2001).

Two essential components define the model: growth by single-gene duplication plus correlated re-wiring. The second rule is inspired in the assumption that novelties derived from changes in regulation patterns will be constrained by the functional properties present in already established interacting networks or subnetworks. Such constraints are likely to be relaxed when new genes are created through duplication. Following available data, the interactions associated with the redundant copies are rapidly lost, whereas new connections emerge rarely through proteome evolution. Estimates of these rates have been used in our analysis. Besides, genome-wide studies on the fate of redundant genes reveals that 50% of duplicated genes lose their function after duplication, whereas the other half experience functional divergence (Nadeau and Sankoff, 1997). Although no functionality is used here, it is interesting to mention that more than 40% of vertices in the proteome model proposed here become disconnected from the main component. If such a disconnection can be related to loss of function, then the fraction of loss genes seems to be close to observed rates.

We derived the rate equations for the evolution of the degree distribution $P(k)$ and its stationary states under some constraints imposed by available data from the analysis of yeast proteome. The rate equation successfully reproduces the scaling law reported from proteome analysis, although the exact exponent is no reproduced due to the strong effects of fluctuations implicit in this type of growth mechanism. Although we concentrated our study in comparing model and data distributions (which are assumed to represent steady states) future extensions should consider the time-dependent behavior of the model as well as possible extensions that would treat the problem of how resident genomes degrade in time (Andersson and Kurland, 1998).

Together with the rules that define the evolution of our model proteome, we introduced characteristic rates that are estimated from available information. The rates of change are of course very important, since they are also responsible for the final connectedness, clustering and sparseness of the graph. What are the factors that tune the rates of edge addition and deletion? One possible source of tuning might be related to the cost of wiring. Additional, functional edges require higher transcription levels and are constrained by different sources of regulation feedbacks. A sparse graph might be a topological blueprint of the underlying optimization process operating at the level of protein wiring. Actually, optimization of graphs has been shown to lead to scale-free networks when both edge density and graph distance are minimized simultaneously (Ferrer i Cancho and Solé, 2001). Since network communication plus low cost leads to heterogeneous maps with scaling properties, it might be the case that the resulting homeostasis characteristic of scale-free nets is actually a byproduct of evolutionary dynamics. Selection would be unnecessary in order to explain the overall patterns displayed by this network architecture: by simply maintaining the system communicated at low average degree $\langle k \rangle$, the multiplicative character of proteome growth spontaneously leads to scaling in the degree. However, not any rates of edge removal allows to obtain scale-free edge distributions, and thus selection might have operated at a more broad level.

Further developments of this model should consider different components of proteome structure and the underlying dynamics of protein–protein interactions. The modular structure of cellular networks (Hartwell et al., 1999; Clarcke and Mittenthal, 2001; Ravasz et al., 2002) or the presence of degeneracy and redundancy (Edelman and Gally, 2001) and its relation with other natural and artificial systems, should be explored. It is

also important to understand the relative role of the three cellular networks (genome, proteome and metabolome) in shaping evolutionary paths. In this context, global analysis of metabolic maps might help expand our approach into more appropriate models including functionality (see for example Ouzounis and Karp, 2000).

The fact that scale-free nets seem so widespread might actually provide a new framework for the study of evolutionary convergence: heterogeneous nets might actually result from optimal searches in high-dimensional parameter spaces. In this context, the proteome map would offer an excellent example of a system where selection, optimization, and tinkering might be at work (Solé et al., 2002a).

Acknowledgements

The authors thank Jay Mittenthal, Ramon Ferrer, Jose Montoya, Stuart Kauffman, Baldo Oliva, and Andy Wuensche for useful discussions. This work has been supported by a grant BFM 2001-2154 and by the Santa Fe Institute (RVS). R.P.-S. acknowledges financial support from the Ministerio de Ciencia y Tecnología (Spain).

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions. Dover, New York.

Albert, R.A., Jeong, H., Barabási, A.-L., 2000. Error and attack tolerance of complex networks. Nature 406, 378–382.

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

Andersson, S.G.E., Kurland, C., 1998. Reductive evolution of resident genomes. Trends Microbiol. 6, 263–268.

Bartel, P.L., Roecklein, J.A., SenGupta, D., Fields, S.A., 1996. A protein linkage map of *Escherichia coli* bacteriophage t7. Nature Genet. 12, 72–77.

Bollobás, B., 1985. Random Graphs. Academic Press, London.

Bornholdt, S., Schuster, H.G., 2002. Handbook of Graphs and Networks: From the Genome to the Internet. Springer, Berlin.

Bray, D., 1995. Protein molecules as computational elements in living cells. Nature 376, 307–312.

Clarcke, B., Mittenthal, J.E., 2001. Modularity and reliability in the organization of organisms. Bull. Math. Biol. 54, 1–20.

Edelman, G.M., Gally, J.A., 2001. Degeneracy and complexity in biological systems. Proc. Natl Acad. Sci. USA 98, 13763–13768.

Fell, D., Wagner, A., 2000. The small world of metabolism. Nature Biotech. 18, 1121.

Ferrer i Cancho, R., Janssen, C., Solé, R.V., 2001a. The topology of technology graphs: small world pattern in electronic circuits. Phys. Rev. E 63, 32767.

Ferrer i Cancho, R., Janssen, C., Solé, R.V., 2001b. The small world of human language. Proc. Roy. Soc. London B 268, 2261–2266.

Ferrer i Cancho, R., Solé, R.V., 2001. Optimization in complex networks. Santa Fe working paper 01-11-068.

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., Transy, C., Legrain, P., 2000. A genomic approach to the hepatitis c virus generates a protein interaction map. Gene 242, 369–379.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., Feldman, M.W., 2002. Evolutionary rate in the protein interaction network. Science 296, 750–752.

Fromont-Racine, M., Rain, J.C., Legrain, P., 1997. Towards a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nature Genet. 16, 277–282.

Gardiner, C.W., 1985. Handbook of Stochastic Methods, 2nd Edition. Springer, Berlin.

Gavin, A., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.

Goh, K., Oh, E., Jeong, H., Kahng, B., Kim, D., 2002. Classification of scale-free networks. Proc. Natl Acad. Sci. USA 99, 12583–12588.

Goodwin, B., 1994. How the Leopard Changed its Spots: The Evolution of Complexity. Weidenfeld & Nicholson, London.

Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W., 1999. From molecular to modular cell biology. Nature 402, C47–C52.

Hazbun, T.R., Fields, S., 2001. Networking proteins in yeast. Proc. Natl Acad. Sci. USA 98, 4277–4278.

Ho, Y., 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae*. Nature 415, 180–183.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., Sakaki, Y., 2000. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl Acad. Sci. USA 97, 1143–1147.

Jeong, H., Mason, S., Barabási, A.L., Oltvai, Z.N., 2001a. Lethality and centrality in protein networks. Nature 411, 41.

Jeong, H., Tombor, B., Albert, R., N.Oltvai, Z., Barabasi, A.-L., 2001b. The large-scale organization of metabolic networks. Nature 407, 651–654.

Kim, J., Krapivsky, P.L., Kahng, B., Redner, S., 2002. Infinite-order percolation and giant fluctuations in a protein interaction network. Phys. Rev. E 66, 055101.

Krapivsky, P.L., Redner, S., Leyvraz, F., 2000. Connectivity of growing random networks. Phys. Rev. Lett. 85, 4629.

Kumar, A., Snyder, M., 2001. Protein complexes take the bait. Nature 415, 123–124.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., 2000. Molecular Cell Biology, 4th Edition. W. H. Freeman, New York.

McCraith, S., Holtzman, T., Moss, B., Fields, S., 2000. Genome-wide analysis of vaccinia virus protein–protein interactions. Proc. Natl Acad. Sci. USA 97, 4879–4884.

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

Monod, J., 1970. Le hasard et la nécessité. Editions du Seuil, Paris.

Montoya, J.M., Solé, R.V., 2002. Small world patters in fodd webs. J. theor. Biol. 214, 405–412.

Nadeau, J.H., Sankoff, D., 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics 147, 1259–1266.

Ohno, S., 1970. Evolution by Gene Duplication. Springer, Berlin.

Ouzounis, C.A., Karp, P.D., 2000. Global properties of the metabolic map of *Escherichia coli*. Genome Res. 10, 568–576.

Pastor-Satorras, R., Vázquez, A., Vespignani, A., 2001. Dynamical and correlation properties of the internet. Phys. Rev. Lett. 87, 258701.

Patthy, L., 1999. Protein Evolution. Blackwell, Oxford.

Podani, J., Oltvai, Z., Jeong, H., Tombor, B., Barabási, A.-L., Szathmáry, E., 2001. Comparable system-level organization of Archaea and Eukaryotes. Nature Genet. 29, 54–56.

Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., Chemama, Y., Labigne, A.S., Legrain, P., 2001. The protein–protein interaction map of *Helicobacter pylori*. Nature 409, 743.

Ravasz, E., Somera, S.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297, 1551–1555.

Ross-Macdonald, P., Coelho, P.S.R., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L., Heldtman, M., Nelson, F.K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G.S., Snyder, M., 1999. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature 402, 413–418.

Rzhetsky, A., Gomez, S.M., 2001. Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. Bioinformatics 17, 988–996.

Solé, R.V., Ferrer, R., Montoya, J.M., Valverde, S., 2002a. Selection, tinkering, and emergence in complex networks. Complexity 8, 20–33.

Solé, R.V., Pastor-Satorras, R., Smith, E., Kepler, T., 2002b. A model of large-scale proteome evolution. Adv. Complex. Systems 5, 43–54.

Stephan, K.A., Hilgetag, C.-C., Burns, G.A.P.C., O'Neill, M.A., Young, M.P., Kötter, R., 2000. Computational analysis of functional connectivity between areas of primate cerebral cortex. Philos. Trans. Roy. Soc. B 355, 111–126.

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

Vázquez, A., Flammini, A., Maritan, A., Vespignani, A., 2003. Modelling of protein interaction networks. Complexus 1, 38–44.

Wagner, A., 2000. Robustness against mutations in genetic networks of yeast. Nature Genet. 24, 355–361.

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

Walhout, A.J.M., Sordella, R., Lu, X.W., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., Vidal, M., 2000. Protein interaction mapping in *c. elegans* using proteins involved in vulval development. Science 287, 116–122.

Watts, D.J., 1999. Small Worlds. Princeton University Press, Princeton.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440–442.

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

Wuchty, S., 2001. Scale-free behavior in protein domain networks. Mol. Biol. Evol. 18, 1694–1702.