Evolving protein interaction networks through gene duplication

Romualdo Pastor-Satorras,[1] Eric Smith,[2] and Ricard V. Solé[1,2,3,*]
[1]*Complex Systems Research Group, FEN*
Universitat Politècnica de Catalunya, Campus Nord B4, 08034 Barcelona, Spain
[2]*Santa Fe Institute, 1399 Hyde Park Road, New Mexico 87501, USA*
[3]*Complex Systems Lab, IMiM-UPF, Dr. Aiguader 80, 08003 Barcelona, Spain*

(Dated: February 15, 2002)

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 links 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, suggesting that the overall topology of the protein maps naturally emerges from the two leading mechanisms considered by the model.

I. 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, 2001). 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 the 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 (see for example Hutchison *et al.* (1999)) 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, networks of genes are also the target of selective forces.

Recent large-scale studies of the global properties of the yeast proteome reinforce the relevance of the network perspective (Gavin *et al.*, 2002; Ho *et al.*, 2002; Jeong *et al.*, 2001*a*; Wagner, 2001*a*). These studies have revealed that the available data from protein-protein interaction networks in the yeast *Saccharomyces cerevisiae* share some unexpected features with other complex networks (Jeong *et al.*, 2001*a*; Wagner, 2001*a*). 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 behaviour, $P(k) \approx k^{-\gamma}$, 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 node has a well-defined neighborhood of "close" nodes) but the minimum distance between any two randomly chosen nodes in the graph is short, a characteristic feature of random graphs (Watts, 1999; Watts & Strogatz, 1998). Scale-free (SF) networks appear to be present in many natural and artificial systems, ranging from technological networks (Albert *et al.*, 2000; Amaral *et al.*, 2000; Ferrer i Cancho *et al.*, 2001*a*; Pastor-Satorras *et al.*, 2001), neural networks (Watts & Strogatz, 1998), metabolic pathways (Fell & Wagner, 2000; Jeong *et al.*, 2001*b*; Podani *et al.*, 2001), and food webs (Montoya & Solé, 2002; Williams *et al.*, 2001) to the human language graph (Ferrer i Cancho *et al.*, 2001*b*). It is remarkable, in particular, that the exponents observed in

*Corresponding author.

Internet, metabolic, and protein networks are very similar. This fact hints towards the presence of a common self-organization principle, 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.*, 2001*a*; Wagner, 2001*a*): (1) the proteome map is a sparse graph, indicating a small average number of links per protein. This observation is also consistent with the study of the global organization of the *E. 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 of links 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 rewiring 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., 2002). 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 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.*, 2002*b*). It is also worth noting the work by Vázquez *et al.* (2001), in which a related model of proteome evolution, showing multifractal connectivity properties, is described and analyzed.

This paper is organized as follows. In Sec. 2 we review the known topological properties of proteome networks, obtained by several authors by analyzing the published proteome maps of the yeast *S. cerevisiae*. In Sec. 3 we describe our model of proteome growth. The main ingredients of the model are protein duplication plus correlated random rewiring. Secs. 4 and 5 are devoted to an analytical study of the model. In Sec. 4 we discuss a mean-field approximation for the evolution of the average connectivity, that will allow us to restrict the range of values of the model's parameters, while Sec. 5 presents a study of the rate equation for the node distribution n_k within an approximation that imposes an uncorrelated rewiring of connections after each node duplication. The solution of the model will show us the limited validity of this sort of approach for the model in question. Our study is completed in Sec. 5 by means of computer simulations. In Sec. 6 we present a discussion of our results. Finally, we inspect in an Appendix the rate equation for the node distribution under correlated rewiring, recovering a similar result as in the uncorrelated case.

II. 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 *C. 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 & Fields, 2001), the statistical patterns used in our study seem to be robust. Recent systematic analyses of protein complexes by means of mass spectrometry provided very similar results, together with a better understanding of the internal organization of protein complexes (Kumar & Snyder, 2001).

From a statistical point of view, protein-protein interaction maps can be viewed as a random network (Bollobás, 1985), in which the nodes represent the proteins and a link between two nodes 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 $W_p = \{p_i\}, (i = 1, ..., N)$ 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 network representation of the protein interactions, shown in Fig. 1(a)[1], reveals a very complex topology, characterized by the presence of several highly connected hubs, while most of the proteins have very few connec-

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

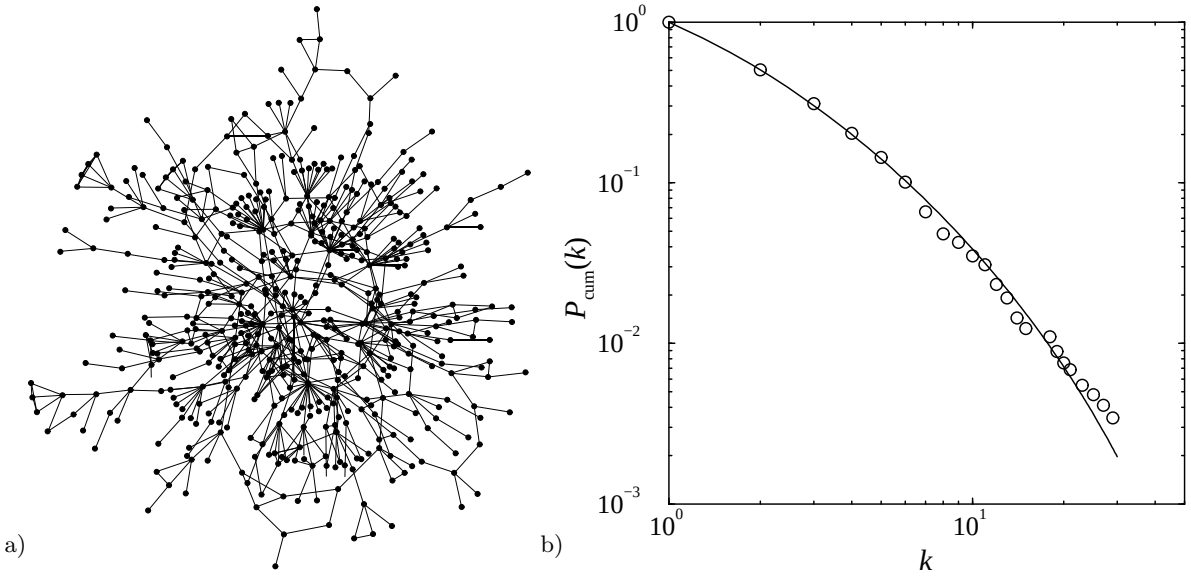

FIG. 1 (a) Topology of a real yeast proteome map obtained from the MIPS database (Mewes *et al.*, 1999). (b) Cumulated degree distribution for the yeast proteome map from Jeong *et al.* (2001*a*). The proteome map is available at the web site http://www.nd.edu/∼networks/database/index.html. The degree 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 $k_c \simeq 15$.

tions. The network topology can be statistically characterized by means of the degree distribution $P(k)$, defined as the probability that any node is connected to exactly k other nodes. The analysis of the protein map from the yeast *S. Cerevisiae*, containing 1870 nodes and 2240 links, corresponding to an average connectivity (average number of links emanating from a node) $\langle k \rangle = 2.40$, shows that the degree distribution can be fitted to a power-law with an exponential cut-off, of the form

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

The estimated values for the yeast are $k_0 \simeq 1$, $\gamma \simeq 2.4$ and $k_c \simeq 20$ (Jeong *et al.*, 2001*a*). 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*a*), who found a power-law behavior with $\gamma \simeq 2.5$ for a relatively smaller protein map (985 nodes with average connectivity $\langle k \rangle = 1.83$). In Fig. 1(b) we have checked this functional dependence on the cumulated degree distribution of the protein map used in Jeong *et al.* (2001*a*) (available at the web site http://www.nd.edu/∼networks/database/index.html). A fit to the 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.* (2001*a*) and Wagner (2001*a*).

An additional observation from Wagner's study of the yeast proteome is the presence of SW properties (Watts & 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 nodes in the network (Bollobás, 1985). Let us consider the adjacency matrix of the giant component, ξ_{ij}^∞ and indicate by $\Gamma_i = \{p_i \,|\, \xi_{ij}^\infty = 1\}$ 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 & Strogatz, 1998). Denoting

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

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

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

where k_i is the connectivity of the i-th 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 node 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:

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

	Wagner (2001*a*)	Map from Jeong *et al.* (2001*a*)	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

TABLE I Comparison between the observed regularities in the yeast proteome reported by Wagner (2001*a*), those calculated from the proteome map used in Jeong *et al.* (2001*a*), the model predictions with $N = 2000$, $\delta = 0.53$ and $\beta = 0.06$ (see Sec. 6), and a random network with the same size and connectivity as the model.

Random graphs, where nodes are randomly connected with a given probability p (Bollobás, 1985), have a clustering coefficient inversely proportional to the network size, $C^{\rm rand} \approx \langle k \rangle / N$, and an average path length proportional to the logarithm of the network size, $\bar{\ell}^{\rm rand} \approx \log N / \log \langle k \rangle$. At the other extreme, regular lattices with only nearest-neighbor connections among units are typically clustered and exhibit a long 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 connectivity and number of nodes, $\bar{\ell} \approx \bar{\ell}^{\rm rand}$.

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

III. 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. After duplication of a single, randomly chosen gene, new connections 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 link in the network. Both processes (creation and deletion of links) might be associated or not to the newly created unit. Four possible combinations are thus allowed in principle: 1) Correlated creation and deletion of links. 2) Correlated creation and uncorrelated deletion of links. 3) Uncorrelated creation and correlated deletion of links. 4) Uncorrelated creation and deletion of links. The rules associated with each variation of the model are summarized in Fig. 2

In this work we will focus in the first variation of the model, in which created and deleted links occur in relation with the newly duplicated node. 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 links is uncorrelated yield results which are in disagreement with the experimental observations in real proteome maps.

The model we will consider is defined by the following rules. We start form a set m_0 of connected nodes, and each time step we performe the following operations:

(i) One node of the graph is selected at random and duplicated

(ii) The links emanating from the newly generated node are removed with probability δ

(iii) New links (not previously present after the duplication step) are created between the new node and all the rest of the nodes with probability α

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 rewiring 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 inteded 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

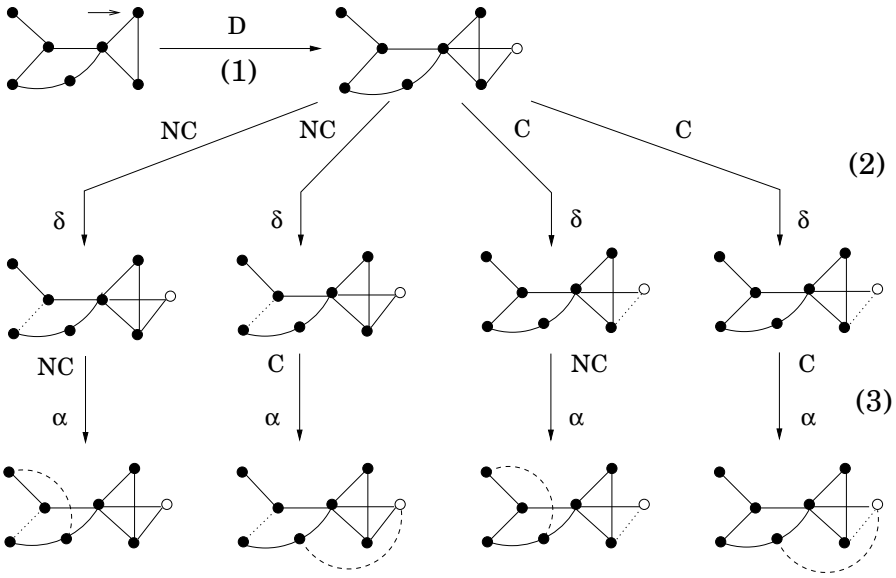

FIG. 2 Rules of proteome growth in the four possible scenarios. First, (1) duplication occurs after randomly selecting a node (small arrow). Then (2) deletion of connections occurs with probability δ. This event can be correlated (C) when the deleted links are connected to the newly generated node or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

due to the duplication and diversification processes. Although through the evolution of genomes genes become non-functional, here we assume that all interactions are functional and thus no pseudogenes are created.

IV. ANALYTICAL STUDY OF THE MODEL: MEAN-FIELD RATE EQUATION FOR THE AVERAGE CONNECTIVITY

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.*, 2001*a*; Wagner, 2001*a*), as well as recent analysis of large-scale perturbation experiments (Wagner, 2001*b*). This can be done for any model with node duplication plus addition/deletion of nodes by considering the discret dynamics of the number or links L_N at a given step N, where N is the number of nodes in the network (see also Vázquez *et al.* (2001)). In general, we can write the evolution equation

$$L_{N+1} = L_N + K_N + \phi_{\rm a}(K_N, L_N) - \phi_{\rm d}(K_N, L_N), \quad (6)$$

where $K_N = 2L_N/N$ indicates average connectivity at the N-th duplication event, and $\phi_{\rm a}$ and $\phi_{\rm d}$ stand for the general rates of addition/deletion of nodes, respectively.

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

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

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

$$\frac{dK_N}{dN} \simeq K_{N+1} - K_N, \quad (8)$$

Eq. (7) can be written

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

whose solution is

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

where $\Gamma = 1 - 2(\alpha + \delta)$ and K_1 is the initial connectivity at $N = 1$. For any value of α and δ this version leads to an increasing connectivity through time. Under this conditions, and in order to have a final sparse graph with a low number of links per protein, we need to consider two possible scenarios. The first would consider 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 connectivity is indicated as K^*, the required number of nodes N^* will be

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

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

Another, more elegant, possibility is to assume that the rate of link creation scales as the inverse of N, i.e. $\alpha = \beta/N$, where $\beta > 0$ is some constant. That is, the rate of addition of new links (the stablishment 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 Wagner (2001*a*). Using this scaling form, the rate equation for K_N reads now

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

The time dependent solution now reads

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

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

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

that yields

$$\begin{aligned} K_N &= N^{1-2\delta}e^{2\beta/N}\left[C + (2\beta)^{2\delta}\Gamma(1-2\delta)\right] \\ &- 2\beta e^{2\beta/N}\sum_{m=0}^{\infty}\frac{(-2\beta/N)^m}{(m+1-2\delta)m!}. \end{aligned}$$

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

$$K_\infty = \lim_{N\to\infty} K_N = \frac{2\beta}{2\delta - 1}. \qquad (15)$$

This is thus consistent with the data analysis by Wagner (2001*a*). Eq. (15) 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

$$\beta = (\delta - 1/2)K_\infty \equiv (\delta - 1/2)\left\langle k\right\rangle. \qquad (16)$$

V. ANALYTICAL STUDY OF THE MODEL: RATE EQUATION FOR THE NODE 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 nodes in the network with exactly k links at time t. Defining our network as the set of numbers $n_k(t)$, we have that the total number of nodes N is given by

$$N = \sum_k n_k, \qquad (17)$$

while the total number of links is given by

$$L = \frac{1}{2}\sum_k k n_k, \qquad (18)$$

since the sum over node connections double-counts links.

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

$$\Pr_{\text{self,dup}}\left[n_k \to n_k + 1\right] = \frac{n_k}{N}(1 - k\delta). \qquad (19)$$

In this expression n_k/N represents the probability of selecting a node of connectivity k and $1-k\delta$ is the probability of preserving all links in the just duplicated node. It is important to note than in Eq. (19) we are ignoring the possibility of deleting more than one link 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 later on that this fact has important consequences when interpreting the results obtained in this Section.

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

$$\Pr_{\text{above,dup}}\left[n_k \to n_k + 1\right] = \frac{n_{k+1}}{N}(k+1)\delta. \qquad (20)$$

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

$$\Pr_{\text{other,dup}}\left[(n_{k-1}, n_k) \to (n_{k-1} - 1, n_k + 1)\right] = \frac{n_{k-1}}{N}(k-1)(1-\delta), \qquad (21)$$

because kn_k is the total number of nodes connected to all nodes of degree k. In Eq. (21) we have corrected for the probability δ that the crucial connecting link was deleted.

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 $N\alpha$ new links to the graph. When this last step is performed with the correlated rule (i.e. adding links from the duplicated node to the rest of the links in the graph), it leads to a nonlocal rate equation for the functions n_k. 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), defering to an Appendix the analysis of the correlated case. We will observe, however, that both cases are doomed to fail, due to the condition $\delta \ll 1$, which is incompatible with the constraint of positive average connectivity.

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

$$\Pr_{\text{add}}\left[(n_k, n_{k+1}) \to (n_k - 1, n_{k+1} + 1)\right] = \frac{n_k}{N} 2\alpha N = \frac{n_k}{N} 2\beta, \tag{22}$$

The probabilities (19), (20), (21), and (22) define the rate equation for the connectivity distribution

$$\frac{dn_k(t)}{dt} = \frac{n_k}{N} + \frac{\delta}{N}\left[(k+1)n_{k+1} - kn_k\right] + \frac{1-\delta}{N}\left[(k-1)n_{k-1} - kn_k\right] + \frac{2\beta}{N}\left[n_{k-1} - n_k\right]. \tag{23}$$

The point to note in Eq. (23) is the first term proportional to n_k/N. This is the unaltered duplication event, which can create a node of degree n_k only by duplicating another such node. It is separated from the rest of link addition probabilities, because for rewired links, there is no correlation between the likelihood that a node of degree k will be created by duplication, and that it will be gained or lost by link addition. Since each time step a new node is added, Eq. (23) satisfies the condition

$$\frac{dN}{dt} = \sum_k \frac{dn_k(t)}{dt} = 1, \tag{24}$$

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. (23), we impose the homogenous condition on the population number

$$n_k(t) = N(t)p_k \simeq tp_k, \tag{25}$$

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

$$(k+1)\delta p_{k+1} - (k + 2\beta)p_k + [(k-1)(1-\delta) + 2\beta]p_{k-1} = 0. \tag{26}$$

Eq. (26) can be solve using the generating functional method (Gardiner, 1985). Let us define the the generating functional

$$\phi(x) = \sum_k x^k p_k. \tag{27}$$

In terms of ϕ, Eq. (26) can be written

$$\left[(1-\delta)x^2 - x + \delta\right]\frac{d\phi(x)}{dx} + 2\beta(x-1)\phi(x) = 0. \tag{28}$$

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

$$\phi(x) = \left(\frac{\delta - x(1-\delta)}{2\delta - 1}\right)^{-2\beta/(1-\delta)}. \tag{29}$$

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

$$\langle k\rangle = \sum_k kp_k \equiv x\frac{d\phi(x)}{dx}\bigg|_{x=1} = \frac{2\beta}{2\delta - 1}, \tag{30}$$

in agreement with the mean-field prediction of Eq. (15).

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

$$p_k = \frac{1}{k!}\phi^{(k)}(0), \tag{31}$$

where $\phi^{(k)}(x)$ is the k-th derivative of $\phi(x)$. Appliying this formula on the function (29), we are led to

$$p_k = \left(\frac{2\delta-1}{\delta}\right)^{2\beta/(1-\delta)} \frac{1}{\Gamma\left(\frac{2\beta}{1-\delta}\right)} \frac{\Gamma\left(\frac{2\beta}{1-\delta}+k\right)}{k!} \left(\frac{\delta}{1-\delta}\right)^{-k}. \tag{32}$$

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

$$p_k \sim (k_0+k)^{-\gamma} e^{-k/k_c}, \tag{33}$$

with

$$\gamma = -k_0 = 1 - \frac{2\beta}{1-\delta}, \qquad k_c = \frac{1}{\ln\left(\frac{\delta}{\delta-1}\right)}. \tag{34}$$

As we observe from Eq. (33), we recover the same functional form experimentally observed in Jeong *et al.* (2001*a*). 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. (34), that is $\gamma \leq 1$. As we will see in the Appendix, the same result holds when we consider the rate equation for the full model, in which the link addition is fully correlated with the new duplicated node. 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 has only meaning for $\delta > 1/2$ (see Eq. (30)). Yet the master equation was defined on the basis of an independent-event approximation that only makes sense for $\delta \ll 1/2$. 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, $\langle k \rangle$.

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). Fortunately, as we will see in the next Section, just in this area a trend from the simulations seems to at least qualitatively meet the value given by the analytic solution.

VI. NUMERICAL RESULTS

The proteome model defined in Section 3 depends effectively on two independent paramenters: the average connectivity of the network $\langle k \rangle$ and the deletion rate of newly created links δ; given these two parameters, the rate β can be computed from Eq. (16). The average connectivity can be estimated from the experimental results from real proteome maps. Examination of Table I yields a value $\langle k \rangle \simeq 2.40$ from the data analyzed in Jeong *et al.* (2001*a*). As a safe estimate, we impose the value $\langle k \rangle = 2.5$ in our model. In Solé *et al.* (2002*b*) the rate δ was roughly estimated from the experimentally calculated ratio of addition and deletion rates in the yeast proteome, α/δ (Wagner, 2001*a*). However, it is clear that this estimate is strongly dependent of the assumed value α/δ. In this work we will consider instead the more general case of a δ-dependent model. In spite of the drawbacks of the analytical study in Sec. 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 connectivity $\langle k \rangle = 2.5$). 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.

Simulations of the model start from a connected ring of $N_0 = 5$ nodes and proceed by interating the rules of the model until the desired network size is achieved. Given the size of the maps analized by Jeong *et al.* (2001*a*), we consider networks with $N = 2 \times 10^3$ nodes. In Fig 3 we plot the values of γ estimated from the 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 non-linear regression of the corresponding degree distribution in the range $k \in [1, 80]$. In this Figure we observe that, apart form 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

$$\delta = 0.562. \tag{35}$$

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

In Figure 4(a) we show the topology of the giant com-

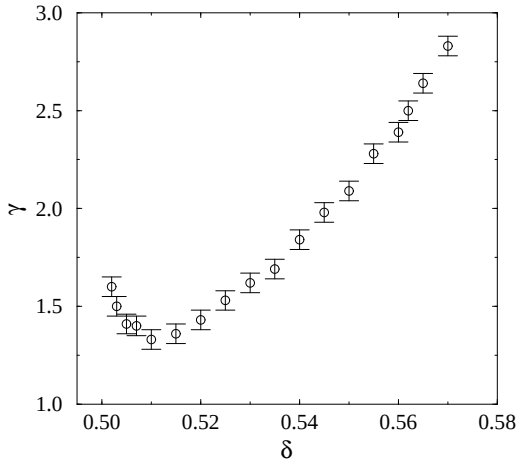

FIG. 3 Degree exponent γ as a function of the deletion rate δ from computer simulations of the proteome model with average connectivity $\langle k \rangle = 2.5$. Network size $N = 2 \times 10^3$. The degree distributions is averaged over 1000 different network realizations.

ponent 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 Figure 1(a); we can appreciate the presence of a few highly connected hubs plus many nodes with a relatively small number of connections. On the other hand, in figure 4(b) we plot the connectivity $P(k)$ obtained for networks of size $N = 2 \times 10^3$, averaged of 10000 realizations. 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), with parameters $\gamma = 2.5 \pm 0.1$ and $k_c \simeq 37$, in fair agreement with the measurements reported by Wagner (2001*a*) and Jeong *et al.* (2001*a*).

We have also computed the SW properties of the model. In Table I 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.*, 2001*a*), and the values corresponding to a random graph with size and 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 Wagner (2001*a*).

VII. DISCUSSION

In this paper a detailed analysis of a model of proteome evolution (Solé *et al.*, 2002*b*) 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 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. 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 & Snyder, 2001).

Two essential components define the model: growth by single gene duplication plus correlated re-wiring. Unequal cross-over is actually known to be the dominant contribution to genome growth and dynamics (Ohno, 1970). 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.

We derived the rate equations for the evolution of the degree distribution $n_k(t)$ and its stationary states under some constraints imposed by available data from the analysis of yeast proteome. Although we concentrated our study in comparing model and data distributions (which are assumed to represent steady states) future analysis should also explore 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 *et al.*, 1998).

Together with the rules that define the evolution of our proteome model, 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 conectedness, clustering and sparseness of the graph. What are the factors that tune the rates of link addition and deletion? One possible source of tuning might be related to the cost of wiring. Additional, functional links, 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 link density and graph distance are minimized simultaneously (Ferrer i Cancho & 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. In such a case, we would reach "robustness for free" as an emergent property.

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 (Clarcke & Mittenthal, 2001; Hartwell *et al.*, 1999) or the presence of degeneracy and redundancy (Edelman & Gally, 2001)

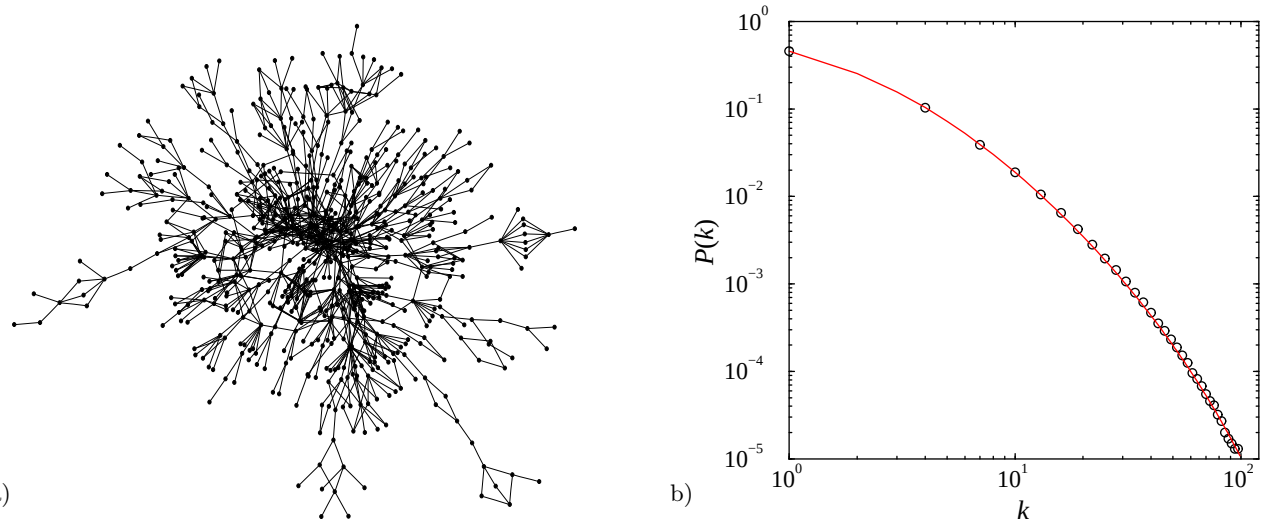

FIG. 4 (a) Topology of the giant component of the map obtained with the proteome model with parameters $\langle k\rangle = 2.5$ and $\delta = 0.565$. Network size $N = 2 \times 10^3$. (b) Degree distribution for the same model, averaged over 10000 different network realizations.

and its relation with other natural and artificial systems, should be explored. 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.*, 2002*a*).

The authors thank Jay Mittenthal, Ramon Ferrer, Jose Montoya, Stuart Kauffman, and Andy Wuensche for useful discussions. This work has been supported by a grant PB97-0693 and by the Santa Fe Institute (RVS). R.P.-S. acknowledges financial support from the Ministerio de Ciencia y Tecnología (Spain).

Appendix

In this Appendix we will consider the rate equation for the node distribution n_k for the full correlated version of the model, in which all the new added links emanate from the duplicated node. While the terms due to node duplication and link removal remain the same (Eqs. (19) to (21)), the correlated nature of the link addition radically changes the form of the addition term (22). In this case, half of the link ends that can modify a node of degree k are still induced by a duplication of some other node of unrelated degree. These ends remain proportionally distributed as in Eq. (22):

$$\Pr_{\text{add,other}}\left[(n_k, n_{k+1}) \to (n_k - 1, n_{k+1} + 1)\right] = \frac{n_k}{N}(\alpha N)\,. \tag{36}$$

One out of every two new link ends, though, remains attached to the duplicated node, potentially shifting its index from the k of pure duplication to some $k + n$. In any such addition scenario, all of the original nodes at degree k remain as they were, and the probability for the new node to add one to the population at degree $k + n$ is

$$\Pr_{\text{add,self}}\left[(n_k, n_{k+n}) \to (n_k, n_{k+n} + 1)\right] = \frac{n_k}{N}\binom{N}{n}\alpha^n(1-\alpha)^{N-n}. \tag{37}$$

Eq. (37) can be broken down into the following components: n_k/N is the probability that the duplicated node has degree k. Given that that node is duplicated, every one of the N pre-existing nodes in the graph is considered as a candidate for a link addition to it, and accepted with probability α. After all such links have been considered, the probability of any configuration with n additions is $\alpha^n(1-\alpha)^{N-n}$, and the number of such configurations possible is the binomial coefficient for n successful additions out of N independent tries.

The check that the correct number of link ends has been added is that

$$\sum_n n\binom{N}{n}\alpha^n(1-\alpha)^{N-n} = \alpha N, \tag{38}$$

which with the (αN) of Eq. (36), gives the $(2\alpha N)$ of the uncorrelated addition model (22).

The probability that a new node of degree k will be created by duplication and no other event is now $(1-\alpha)^N$ times that in the uncorrelated case. At linear order, this is still the dominant term modified by deletions, so the $-\delta k n_k/N$ will remain the same with correlations as in Eq. (23). However, it is no longer correct to give all the addition weight to n_k; rather it must be binomially distributed over all the contributing probabilities at degrees $k-n$. Making this replacement, the rate equation for the node distribution in the correlated model takes the form

$$\begin{aligned}\frac{dn_k(t)}{dt} &= \frac{1}{N}\sum_{n=0}^{k} n_{k-n}\binom{N}{n}\alpha^n(1-\alpha)^{N-n} + \frac{\delta}{N}\left[(k+1)n_{k+1} - kn_k\right] \\ &+ \frac{1-\delta}{N}\left[(k-1)n_{k-1} - kn_k\right] + \frac{\beta}{N}\left[n_{k-1} - n_k\right]. \end{aligned} \tag{39}$$

Imposing again the homogenous condition Eq. (25) we can solve Eq. (39) using the generating functional defined in Eq. (27). The only new term involves the binomial factor in Eq. (39). Multiplying the binomial term by x^k and summing over $k \in 0, \ldots, N$, one can change the order of summation, to get

$$\sum_{k=0}^{N}\sum_{n=0}^{k} x^{k-n} p_{k-n}\binom{N}{n}(\alpha x)^n(1-\alpha)^{N-n} = \sum_{n=0}^{N}\binom{N}{n}(\alpha x)^n(1-\alpha)^{N-n}\sum_{m=0}^{N-n} x^m p_m. \tag{40}$$

Using Stirling's formula to approximate the factorials in the binomial distribution, solving for the stationary n in the resulting exponential function, and using the fact that $\alpha = \beta/N$, it then follows that, at large N, all terms give non-negligible weight by the binomial cluster around a finite n, so that the moment-generating sums satisfy

$$\sum_{m=0}^{N-n} x^m p_m \approx \sum_{m=0}^{N} x^m p_m \approx \phi(x). \tag{41}$$

Eq. (40) then factors, with the weight terms summing to

$$\sum_{n=0}^{N}\binom{N}{n}(\alpha x)^n(1-\alpha)^{N-n} = \left[1-(1-x)\frac{\beta}{N}\right]^N \overset{N\to\infty}{\longrightarrow} \exp\left[-\beta\,(1-x)\right]. \tag{42}$$

Using Eqs. (40) and (42) in the rate equation, yields the final equation for the generating functional ϕ

$$\left[\delta - x\,(1-\delta)\right]\frac{d\phi}{dx} = \beta\psi\,(x)\,\phi\,(x)\,, \tag{43}$$

where the auxiliary function $\psi(x)$ has the form, valid for large N,

$$\psi\,(x) = \frac{1+\beta(1-x) - \exp\left[-\beta\,(1-x)\right]}{\beta(1-x)}. \tag{44}$$

Given the general relation (43) for ϕ, it is possible to write a closed-form recursion relation for the moments p_k. It remains a differential relation, but gives some ability to recover the parameters of the asymptotic form for large k.

The first step is to recall Eq. (31), relating the degree distribution to the derivatives of ϕ. One then changes variables from x to the natural variable $y \equiv \beta\,(1-x)$, in terms of which the large-N form of $\psi\,(y)$ is

$$\psi\,(y) = \frac{1+y-e^{-y}}{y}. \tag{45}$$

Defining two constant combinations of parameters

$$\upsilon \equiv \frac{2\beta}{1-\delta}, \tag{46}$$

and

$$\omega \equiv \beta\left(\frac{2\delta-1}{1-\delta}\right), \tag{47}$$

Eq. (43) can be written in the simple form

$$-\frac{1}{\beta}\frac{d\phi}{dx}\bigg|_{x=0} = \frac{d\phi}{dy}\bigg|_{y=\beta} = -\frac{\upsilon}{2}\frac{\psi(y)}{\omega+y}\phi(y)\bigg|_{y=\beta} \equiv -\chi(y)\,\phi(y)|_{y=\beta}. \tag{48}$$

It follows from the definition (27) that

$$\langle k\rangle = \beta\chi(y)|_{y\to 0} = \frac{\beta\upsilon}{\omega} = \frac{2\beta}{2\delta-1}, \tag{49}$$

for either correlated or uncorrelated additions, and from the definition (31) that

$$\frac{p_1}{p_0} = \beta\chi(y)|_{y=\beta}. \tag{50}$$

Further, by repeatedly applying Eq. (48), one can obtain an expression for p_k/p_0 in terms of elementary functions, which is not possible for ϕ itself:

$$\frac{p_k}{p_0} = \frac{\beta^k}{k!}\left(\chi(y) - \frac{d}{dy}\right)^{k-1}\chi(y)\bigg|_{y=\beta}. \tag{51}$$

Eq. (51) can be made simpler by regarding υ and ω as constant parameters, and $y=\beta$ as the function argument, with differentiation denoted as $\partial/\partial\beta$ in place of d/dy. In this notation

$$p_0 = \phi(\beta), \tag{52}$$

Eq. (48) becomes

$$\frac{\partial\phi}{\partial\beta} = -\chi(\beta)\,\phi, \tag{53}$$

and the definition (31) with repeated application of Eq. (53) gives

$$p_{k+1} = \frac{[k-\partial/\partial\log\beta]}{[k+1]}p_k. \tag{54}$$

However, it is not p_1 itself that provides a simple initial condition for iteration, but rather $p_1/p_0 = \beta\chi(\beta)$, by Eq. (50). For this reason it is convenient to define

$$\hat{p}_k \equiv \frac{p_k}{p_0}. \tag{55}$$

Eq. (54) is then readily extended with Eq. (53), to produce the desired recursion relation

$$\hat{p}_{k+1} = \frac{[k+\beta\chi(\beta)-\partial/\partial\log\beta]}{[k+1]}\hat{p}_k. \tag{56}$$

The surprising feature of Eq. (56) is that, acting on an ansatz for the asymptotic degree distribution, it provides tractable constraints on the power law and exponential cutoff. The purpose of the exact solution above was to check that this is not an indication that the recursion (56) is in error, but rather a problem with asymptotic expansion of ansätze in powers of large k.

One asks when the recursion (56) is compatible with the leading asymptotic form for $\hat{p}_k$

$$\hat{p}_k = (k_0+k)^{-\gamma}e^{-k/k_c}. \tag{57}$$

In general, k_0, γ, and $1/k_c$ are allowed to be functions of β explicitly, as well as of the constant parameters υ and ω.

For this application, Eq. (56) is more conveniently written as

$$\frac{\hat{p}_{k+1}}{\hat{p}_k} = \frac{[k+\hat{p}_1-\partial\log\hat{p}_k/\partial\log\beta]}{[k+1]}. \tag{58}$$

The lefthand side of Eq. (58) in this ansatz takes the form

$$\frac{\hat{p}_{k+1}}{\hat{p}_k} = \left(1 + \frac{1}{k_0 + k}\right)^{-\gamma} e^{-1/k_c}, \tag{59}$$

which can be expanded as a power series in $1/k$ at large k.

The righthand side of Eq. (58) becomes

$$\frac{[k + \hat{p}_1 - \partial \log \hat{p}_k / \partial \log \beta]}{[k+1]} = \frac{1}{[k+1]} \left[k \left(1 + \frac{\partial (1/k_c)}{\partial \log \beta}\right) + \frac{\partial \gamma}{\partial \log \beta} \log (k_0 + k) + \hat{p}_1 + \frac{\partial k_0}{\partial \log \beta} \frac{\gamma}{k_0 + k} \right]. \tag{60}$$

Since Eq. (59) is a pure power law in $1/k$, the two sides cannot be matched unless

$$\frac{\partial \gamma}{\partial \log \beta} = 0. \tag{61}$$

If Eq. (61) is satisfied, the $\mathcal{O}\left(k^0\right)$ term requires that

$$\left(1 + \frac{\partial (1/k_c)}{\partial \log \beta}\right) = e^{-1/k_c}. \tag{62}$$

The $\mathcal{O}\left(k^{-1}\right)$ term then requires

$$\gamma = 1 - \hat{p}_1 e^{1/k_c}. \tag{63}$$

Eq. (61) is equivalent to the condition $\partial \log (1-\gamma) / \partial \log \beta = 0$, which with Eq. (62) and Eq. (63) evaluates to

$$e^{-1/k_c} = 1 - \frac{\partial \log \hat{p}_1}{\partial \log \beta}. \tag{64}$$

For correlated addition,

$$\hat{p}_1 = \frac{v}{2} \frac{1 + \beta - e^{-\beta}}{\omega + \beta} > 0. \tag{65}$$

For $0 \leq \omega < 1$, it is possible to show that there are solutions to $\partial \hat{p}_1 / \partial \beta = 0$, implying $k_c \to \infty$. For all ω, at large β

$$\frac{\partial \log \hat{p}_1}{\partial \log \beta} \to \frac{\omega}{\omega + \beta} = \frac{2\delta - 1}{\delta}, \tag{66}$$

which is the exact solution in the uncorrelated case. Thus k_c can be made as large as desired in either case, by taking $\delta \to (1/2)^+$.

However, the power law given in Eq. (63) is always strictly less than one, whereas the value from simulations typically is $\gamma > 1$. The only constraint on parameter values from the consistency condition (61) is given through Eq. (64), that the derivative of $\hat{p}_1$ with β be positive. In parameters, this is satisfied when

$$(1 + \omega + \beta) e^{-(\omega + \beta)} \geq (1 - \omega) e^{-\omega}. \tag{67}$$

Presumably, then, the results of this Appendix can be interpreted as follows. For any parameters satisfying Eq. (67), there is a best-fit approximation to the degree distribution with the large-k asymptotic form (57) So the ansatz itself is no restriction. The result of fitting to that ansatz then always has power $\gamma < 1$, for either correlated or uncorrelated addition. Given that the approximations in extracting the degree distribution from the generating functional were well controlled, the problem must lie in the master equation itself.

In fact, the root of the problem is that in order to ensure a finite average degree, the value of δ must be constrained to be larger than 1/2. All the approximations of isolated, single events, made in constructing the equation itself are not very good for such large δ. In particular, the corrected probability $(1 - k\delta) \leq 0$ for all $k \geq 2$ in Eq. (23). The correct coefficient for this term would have been $(1-\delta)^k$, but then a nonlocal master term would have been required for cascades down from all higher $k + n$ in the same line, rather than just from $k + 1$.

A similar problem afflicts all of the addition terms, and addition/deletion interactions. This is particularly serious for correlated additions, where the relation described in Eq. (39) becomes completely wrong, because weights have to be kept for all combinations of addition and deletion terms, rather than treated independently. Further work in necessary in order to clarify this aspects of the rate equation approach for the correlated model.

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of mathematical functions.* New York: Dover.

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. and 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.* London: Academic Press.

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. (2001*a*). 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. (2001*b*). The small world of human language. *Procs. 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.

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 Berlin: Springer.

Gavin, A.-C. et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.

Goodwin, B. (2001). *How the leopard changed its spots: The evolution of complexity.* Princeton: Princeton University Press.

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. et al. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae.* *Nature* **415**, 180-183.

Hutchison III, C. A., Peterson, S. N., Gill, S. R., Cline, R. T., White, O., Fraser, C. M., Smith, H. O., Craig Venter, J. (1999). Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science* **286**, 2165-2169.

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. (2001*a*). Lethality and centrality in protein networks. *Nature,* **411**, 41.

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

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 New York: W. H. Freeman.

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.* Paris: Editions du Seuil.

Montoya, J. M. & Solé, R. V. (2002). Small world patters in fodd webs. J. Theor. Biol. *J. Theor. Biol.* (in press).

Ohno, S. (1970). *Evolution by gene duplication.* Berlin: Springer.

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.* Oxford: Blackwell.

Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabási, A.-L. & Szathmáry, E. (2001). Comparable system-level organization of Archaea and Eukaryotes. *Nature Genetics,* **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.

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.

Solé, R. V., Ferrer, R., Montoya, J. M. & Valverde, S. (2002*a*). Selection, tinkering, and emergence in complex networks. *Complexity* (in press).

Solé, R. V., Pastor-Satorras, R., Smith, E. D. & Kepler, T. (2002*b*). A model of large-scale proteome evolution. *Adv. Complex. Syst.* (in press).

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. (2001). Modelling of protein interaction networks. cond-mat/0108043.

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

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

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

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: Princeton University Press.

Watts, D. J. & Strogatz, S. H. (1998). Colective 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.