ComPlexUs

ORIGINAL RESEARCH PAPER

ComPlexUs 2003;1:38–44; DOI:10.1159/000067642
Received: March 25, 2002
Accepted after revision: October 22, 2002

Modeling of Protein Interaction Networks

Alexei Vázquez[a] Alessandro Flammini[a]
Amos Maritan[a,b] Alessandro Vespignani[b]

[a] International School for Advanced Studies and INFM and
[b] The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

Key Words
Protein interaction network · Duplication · Divergence · Multifractal ·
PACS No. 89.75.-k · 87.15.Kg · 87.23.Kg

Abstract
We introduce a graph-generating model aimed at representing the evolution of protein interaction networks. The model is based on the hypothesis of evolution by duplication and divergence of the genes which produce proteins. The obtained graphs have multifractal properties recovering the absence of a characteristic connectivity as found in real data of protein interaction networks. The error tolerance of the model to random or targeted damage is in very good agreement with the behavior obtained in real protein network analyses. The proposed model is a first step in the identification of the evolutionary dynamics leading to the development of protein functions and interactions.

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com
KARGER

1424–8492/03/0011–0038
$19.50/0
Accessible online at:
www.karger.com/cpu

Amos Maritan
SISSA, Via Beirut 2–4
I–34014 Trieste (Italy)
Tel. + 39 040 2240462, Fax + 39 040 3787528
E-Mail maritan@sissa.it

Synopsis

Proteins interact with one another. Biology is complex. Hence, one might expect that even for a simple organism, the detailed map of the interactions between its various proteins would be a tangled web of overwhelming complexity, more or less random in its appearance. Yet as researchers have recently learned, networks of protein-protein interactions are very far from being random; instead, they possess subtle but significant order. In pioneering experiments, Uetz et al. [1] have mapped out 2,238 pair-wise interactions between 1,825 proteins in the single-celled yeast *Saccharomyces cerevisae* – also known as baker's or brewer's yeast. As it turns out, this network reveals noteworthy mathematical regularities and shares important topological features with other complex networks, including webs of social interaction, the World Wide Web and the Internet.

Like these other networks, the yeast protein network has the 'small world' property – following along links in the network, it requires only a handful of steps to go from any one protein to any other. Another similarity is the manner in which the links are shared out among the various proteins: empirically, the probability that a protein interacts with k other proteins follows a power-law distribution, $P(k) \sim k^{-\gamma}$, with $\gamma \cong 2.5$. Networks of this sort are called 'scale-free', and the average number of links $\langle k \rangle$ offers little insight into the real network topology, which is highly heterogeneous. The 'connectivity' of the various nodes varies considerably: while many nodes have only 1 or a few links, a handful of 'super-connected' hubs have very many.

What is the origin of this architecture? Since scale-free networks arise in diverse settings, it seems likely that some general process might be at work, and this appears to be the case. A crucial insight is that real-world networks have grown to be what they are. Suppose a network starts out small and grows through the progressive addition of new elements, and that each new element, upon entering the network, establishes links at random to a few of the older elements. As Barabási and Albert [5] showed a few years ago, the scale-free architecture

The complete genome sequencing gives for the first time the means to analyze organisms on a genomic scale. This implies the understanding of the role of a huge number of gene products and their interactions. For instance, it becomes a fundamental task to assign functions to uncharacterized proteins, traditionally identified on the basis of their biochemical role. The functional assignment has progressed considerably by partnering proteins of similar functions, leading eventually to the drawing of protein interaction networks (PINs). This has been accomplished in the case of the yeast *Saccharomyces cerevisiae*, where two hybrid analyses and biochemical protein interaction data have been used to generate a web-like view of the protein-protein interaction network [1]. The topology of the obtained graph has been recently studied in order to identify and characterize its intricate architecture [2, 3]. Surprisingly, the PIN resulted in a very highly heterogeneous network with scale-free (SF) connectivity properties.

SF networks differ from regular (local or non-local) networks in the small average link distance between any two nodes (small-world property) [4] and a statistically significant probability that nodes possess a large number of connections k compared to the average connectivity $\langle k \rangle$ [5]. This reflects in a power-law connectivity distribution $P(k) \sim k^{-\gamma}$ for the probability that a node has k links pointing to other nodes. These topological properties, which appear to be realized in many natural and technological networks [6], are due to the interplay of the network growth and the 'preferential attachment' rule. In other words, new appearing nodes have a higher probability to get connected to another node already counting a large number of links. These ingredients encompassed in the Barabaśi and Albert model [5] are present in all the successive SF network models [7–9] and are fundamental for the spontaneous evolution of networks in an SF architecture. From this perspective it is natural to ask about the microscopic process that drives the placement of nodes and links in the case of the protein network.

In the present article we study a growing network model whose microscopic mechanisms are inspired by the duplication and the functional complementation of genes [2, 10, 11]. In this evolutive model all proteins in a family evolved from a common ancestor through gene duplications and mutations (divergence), and the protein network is the blueprint of the entire history of the genome evolution. This duplication-divergence (DD) model is analyzed by analytical calculations and numerical simulations in order to characterize the topological and large-scale properties of the generated networks. We find that the networks show the absence of any characteristic connectivity and exhibit a connectivity distribution with multifractal properties. The n-th moment of the distribution behaves as a power law of the network size N with an exponent that has a nonlinear dependency with n. In addition, the evolution of all moments with N crosses from a divergent behavior to a finite asymptotic value at a given value of the mutation parameters. The generated networks can be directly compared with the *S. cerevisiae* PIN data reported on the Biotech website [12], composed of 1,825 proteins (nodes) connected by 2,238 identified interactions (links). In agreement with the PIN analysis we find that the connectivity distribution can be approximately fitted by a single apparent power-law behavior, and the model's parameter can be optimized in order to quantitatively reproduce magnitudes such as the average connectivity or the clustering coefficient. Further, we compare the simulated tolerance of the DD network to random deletion of individual nodes with that obtained by deleting the most connected ones. While the DD network proves to be fragile in the latter case, it is extremely resistant to random damages, in agreement with the behavior recently observed for the PIN [3]. The analogous topological properties found in the proposed model and the experimental PIN represent a first step towards a possible understanding of protein-protein interactions in terms of gene evolution.

Proteins are divided in families according to their sequence and functional similarities [13, 14]. The existence of these families can be explained using the evolutive emerges naturally from this algorithm if the placement of links follows a rule of 'preferential attachment', that is, if the probability of linking to a node grows in direct proportion to the number of links that node has already.

For a network such as the World Wide Web it is easy to see how this mechanism might come into play. The creator of a new web page naturally provides some links to other pages, and the links chosen will to some extent reflect the visibility of these sites; one has to at least know about a site in order to link to it. People tend to provide links to popular sites that already have a large number of links pointing to them (such as Yahoo, Amazon or Google), rather than to more obscure sites. Hence, preferential attachment may well explain the scale-free character of the World Wide Web.

What about protein networks? It is less clear how preferential attachment might play a role here, and consequently, the origin of the scale-free architecture is more mysterious. But cellular biochemistry has emerged through a long history of biological evolution, and Vázquez et al. show how evolution can produce scale-free networks. They explore a model for the evolution of protein networks that accurately reproduces the topological features seen in the yeast *S. cerevisae*.

As Vázquez et al. point out, proteins fall into families according to similarities in their amino-acid sequences and functions, and it is natural to suppose that such proteins have all evolved from a common ancestor. A favored hypothesis views such evolution as taking place through a sequence of gene duplications – a relatively frequent occurrence during cell reproduction. Following each duplication, the two resulting genes are identical for the moment, and naturally lead to the production of identical proteins. But between duplications, random genetic mutations will lead to a slow divergence of the genes and their associated proteins. This repetitive, two-stage process can be captured in a relatively simple model for the growth of a protein interaction network – the 'duplication-divergence model'.

hypothesis that all proteins in a family evolved from a common ancestor [15]. This evolution is thought to take place through gene or entire genome duplications, resulting in redundant genes. After the duplication, redundant genes diverge and evolve to perform different biological functions. According to the classic model [15], after duplication the duplicate genes have fully overlapping functions. Later on, one of the copies may either become nonfunctional due to degenerative mutations or it can acquire a novel beneficial function and become preserved by natural selection. In a more recent framework [10, 11] it is proposed that both duplicate genes are subject to degenerative mutations and lose some functions, but jointly retain the full set of functions present in the ancestral gene. The outcome of this evolution results in complex PINs with physical, genetic, and biochemical interactions among them. In this article we will restrict our analysis to the physical interactions, and from now on use the term 'interaction' to denote a physical interaction. There is a large number of transient protein-protein interactions that controls and regulates almost all cellular processes. These transient interactions can be detected using the two-hybrid experiments [16]. Two-hybrid experiments are known to be prone to false positives, and it is hard to establish to what extent protein-protein interaction networks can be considered complete and error-free [17]. Nevertheless, the statistical properties of the PIN obtained by different groups are essentially identical, as observed by Yook et al. [18].

The evolution of the PIN can be translated into a growing network model. We can consider each node of the network as the protein expressed by a gene. After gene duplication, both expressed proteins will have the same interactions. This corresponds to the addition of a new node in the network with links pointing to the neighbors of its ancestor. Moreover, if the ancestor is a self-interacting protein, the copy will have also an interaction with it [2]. Eventually, some of the common links will be removed because of the divergence process. We can formalize this process by defining an evolving network in which, at each time step, a node is added according to the following rules:

Duplication: A node i is selected at random. A new node i', with a link to all the neighbors of i, is created. With probability p a link between i and i' is established.

Divergence: For each of the nodes j linked to i and i' we choose randomly one of the two links (i, j) or (i', j) and remove it with probability q.

p is a parameter that models the creation of an interaction between the duplicates of a self-interacting protein and its possible loss due to the divergence of the duplicates. The other parameter q represents the loss of interactions between the duplicates and their neighbors due to the divergence of the duplicates. Our purpose is to provide a minimal model that captures the main effects of the duplication and divergence mechanisms in the PIN topology. Thus p and q take into account fine details in some effective way.

For practical purposes, the DD algorithm starts with two connected nodes and repeats the duplication-divergence rules N times. Since genome evolution analysis [2, 19] supports the idea that the divergence of duplicate genes takes place shortly after the duplication, we can assume that the divergence process always occurs before any new duplication takes place; i.e. there is a time scale separation between duplication and mutation rates. This allows us to consider the number of nodes in the network, N, as a measure of time (in arbitrary units). It is worth remarking that the algorithm does not include the creations of new links; i.e. the developing of new interactions between gene products. This process has been argued to have a probability relatively smaller than the divergence one [2]. However, we found that introducing a probability in the DD algorithm to develop new random connections does not change the network topology.

Another assumption is the one-to-one relation between genes and proteins. It is possible that a gene expresses more than 1 protein. This would mean that a set of proteins, those expressed by the same gene, will have a correlated evolution. In any case, the number of proteins expressed by one

In this model, each node in the network represents a protein that is expressed by a gene, and the network grows as follows. At each time step, one selects a random node – call it node i – and carries out two steps in sequence. First comes duplication. In association with node i, a new node i' enters the network and is linked to the same nodes to which i is linked. This reflects the idea that the new protein, the result of duplication, is identical to the old protein; hence, it interacts with other proteins in the very same way. Also, with some probability p, a link is added between i and i' to account for the possibility that these two (identical) proteins also interact. Next comes divergence. Mutations in the genes associated with i and i' will gradually produce differences in these proteins, altering their interactions. The model accounts for this divergence by considering in turn all the proteins j to which i and i' are linked, selecting one of these at random, and removing it with probability q.

As Vázquez et al. readily acknowledge, this model leaves aside many finer details of the genetic evolution that lies behind the duplication and divergence process. The model aims only to capture the most basic factors affecting the topological evolution of the network. New proteins enter the network following duplication and, because of subsequent mutations, carry with them some but not all of the interaction links of the protein from which they sprouted. Starting with a small seed network (such as two proteins linked to one another) and running the growth procedure many times, one produces a large, complex network. The important question: Do such networks resemble the real protein interaction networks of biology?

One quantity to explore is the distribution of nodes according to their connectivity. Using a computer, Vázquez et al. ran the model 100 times, on each occasion stopping when the number of nodes reached $N = 1{,}825$ – the number of proteins in the available data for the yeast *S. cerevisae.* (As the model follows a process of random growth, the network turns out different in each run; hence, it takes many runs to reveal the average behavior behind the statis-

gene is not comparable with the network size, hence it is expected that a more general model allowing for this correlated evolution will exhibit the same qualitative features.

In order to provide a general analytical understanding of the model, we use a mean-field approach for the moments distribution behavior. Let us define $\langle k \rangle_N$ as the average connectivity of the network with N nodes. After a duplication event $N \longrightarrow N+1$ the average connectivity is given by

$$\langle k \rangle_{N+1} = \frac{(N)\langle k \rangle_N + 2p + (1-2q)\langle k \rangle_N}{N+1} \quad (1)$$

On average, there will be a gain proportional to $2p$ because of the interaction between duplicates, to $\langle k \rangle$ because of duplication, and a loss proportional to $2q\langle k \rangle_N$ due to the divergence process. For large N, taking the continuum limit, we obtain a differential equation for $\langle k \rangle$. For $q > 1/2$, $\langle k \rangle$ grows with N but saturates to the stationary value $k_\infty = 2p/(2q-1) + \mathcal{O}(N^{1-2q})$. By contrast, for $q < 1/2$, $\langle k \rangle$ grows with N as N^{1-2q}. At $q = q_1 = 1/2$ there is a dramatic change of behavior in the large scale connectivity properties. Analogous equations can be written for higher order moments $\langle k^n \rangle$, and for all n we find a value q_n at which the moments cross from a divergent behavior to a finite value for $N \longrightarrow \infty$. More interestingly, it is possible to write the generalized exponents $\sigma_n(q)$ characterizing the moments divergence as $\langle k^n \rangle \sim N^{\sigma n(q)}$. The lengthy calculation that will be reported elsewhere [23] gives the mean-field estimate.

$$\sigma_n(q) = n(1-q) - 2 + 2(1-q/2)^n. \quad (2)$$

The nonlinear behavior with n is indicative of a multifractal connectivity distribution. In order to support the analytical calculations, we performed numerical simulations generating DD networks with a size ranging from $N = 10^3$ to $N = 10^6$. In figure 1 we show the generalized exponents $\sigma_n(q)$ as a function of the divergence parameter q. As predicted by the analytical calculations, $\sigma_n = 0$ at a critical value q_n. The general phase diagrams obtained is in good agreement with the mean-field predictions and the multifractal picture.

Noticeably, multifractal features are present also in a recently introduced model of growing networks [20] where, in analogy with the duplication process, newly added nodes inherit the network connectivity properties from parent nodes. Thus multifractality appears related to local inheritance mechanisms. Multifractal distributions have a rich scaling structure, where the SF behavior is characterized by a continuum of exponents. This behavior is, however, contrary to the usual exponentially bounded distribution and it is interesting to understand why, in the DD model, we generate the large connectivity fluctuation needed for SF behavior. A given node with connectivity k receives, with probability k/N, a new link if one of its neighbors is chosen in the duplication processes. In this case the newly added node will establish a new link to it with a probability of $1-q$. Hence, the probability that the degree of a node increases by one is

$$\omega_{kN} \sim (1-q)k/N, \quad (3)$$

where we have neglected the constant contribution given by the self-interaction probability. This shows that even if evolutionary rules of the DD model are local, they introduce an effective linear preferential attachment known to be at the origin of SF connectivity distribution [21, 24, 25]. However, because the link deletion of duplicate nodes introduces additional heterogeneity to the problem we obtain a multifractal behavior.

The peculiarities of the duplication and divergence process manifest quantitatively in other features characterizing the topology of the network, e.g. the tendency to generate biconnected triplets and quadruples of nodes. These are sets of nodes connected by a simple cycle of links, thus forming a triangle or a square. In the DD model, triangle formation is a pronounced effect since with probability $p(1-q)$, the duplicating genes and any neighbor of the parent gene will form a new triangle. Analogously, duplicating genes and any couple of neighbors of the parent gene will form a new square with the probability $(1-q)^2$. An indication of triangle formation in networks tical fluctuations.) The model has two adjustable parameters p and q, and, as figure 2 shows, it generates networks that closely resemble the real network in yeast if one chooses $p = 0.1$ and $q = 0.7$.

For both the model and for the real data, the figure shows Zipf plots – the logarithm of the number of links k in a node versus the logarithm of the node's rank r (the most highly-connected node having rank 1, the next most highly-connected having rank 2, and so on). For intermediate values of k, both curves are approximately linear, reflecting a power-law or scale-free relationship between connectivity k and rank r. This is not quite the probability distribution of nodes according to their connectivity, but it is easy to show that these two quantities are closely related – and if one is a power-law relationship, then so is the other. Hence, the Zipf plots imply that the distribution of nodes by connectivity is also scale-free – the model captures this aspect of the real network. It also turns out that any single network generated by the model also follows this scale-free pattern, with the exponent $\gamma \simeq 2.5$, as was found empirically for the yeast protein network.

These results appear to reflect a mechanism of preferential attachment that is effective in the evolution of protein networks. In the duplication-divergence model, as Vázquez et al. point out, the node to be duplicated at each time step is chosen at random over the full network; hence highly linked nodes have a better chance of being a neighbor of the selected node than do less connected nodes. Indeed, the chance of a node being the neighbor of the node selected for duplication increases in direct proportion to its connectivity k, since highly connected nodes have more neighbors. Consequently, more highly connected nodes have a greater chance of receiving one of the new links created in the duplication process, and preferential attachment sneaks into the model in a subtle way.

But the duplication-divergence model also has some richer features that distinguish it from most of the earlier models of network growth based on the preferential attachment idea. Power-law distributions are closely associated with fractals – highly

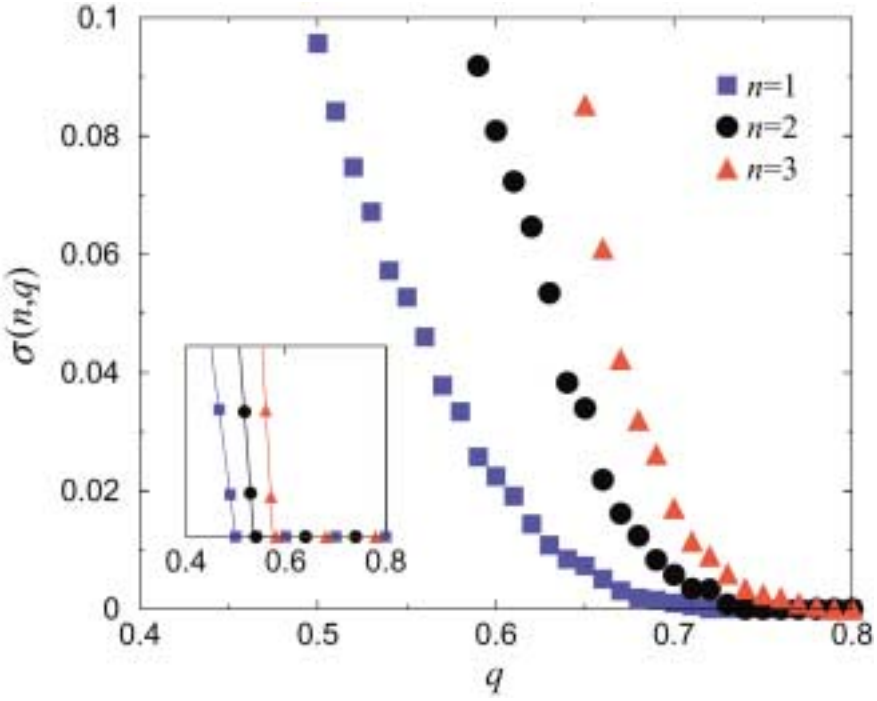

Fig. 1. The exponent $\sigma_n(q)$ as a function of q for different values of n. The symbols were obtained from numerical simulations of the model. The moments $\langle k^n \rangle$ were computed as a function of N in networks with a size ranging from $N = 10^3$ to $N = 10^6$. The exponents $\sigma_n(q)$ are obtained from the power-law fit of the plot $\langle k^n \rangle$ vs. N. In the inset we show the corresponding mean-field behavior, as obtained by equation 2, which is in qualitative agreement with the numerical results.

is given by the clustering coefficient $C_\triangle = 3N_\triangle/N_\Lambda$ [26], where $N_\triangle$ is the number of biconnected triplets (triangles) and N_Λ is the total number of simply connected triplets. Similarly, it is possible to define the square coefficient $C_\square = 4N_\square/N_{II}$, with $N_\square$ the number of squares in the network and N_{II} the number of simply connected quadruples. By measuring these quantities in the yeast *S. cerevisiae* in the Biotech PIN [12], we obtained $C_\triangle = 0.23$ and $C_\square = 0.11$. These values are one order of magnitude larger than those obtained for a SF random graph and other growing network models, for which it has been shown that the clustering coefficient is algebraically decaying with the network size [21]. By contrast, the DD model shows clustering coefficients saturating at a finite value, and it is possible to tune the parameters p and q in order to recover the real data estimates, keeping the average degree as that of the PIN $\langle k \rangle \approx 2.4$. A reasonable agreement with the values obtained for the real PIN is found when $p \simeq 0.1$ and $q \simeq 0.7$, which yields networks with $C_\triangle = 0.10(5)$ and $C_\square = 0.10(2)$. The value of p obtained in this way is close to the fraction of self-interacting proteins reported for the PIN (0.04) [1]. Thus, considering that p is an effective parameter that takes into account self-interactions but that may also include other effects, the agreement is very good. Noticeably, for these values of the parameters the DD model generates networks where other quantities are in good agreement with those obtained from experimental data. A pictorial representation of this agreement is provided in figure 2, where we compare the Zipf plot of the connectivity obtained from 10^3 realizations of the DD model with optimized p and q and that of the yeast *S. cerevisiae* PIN. The DD networks are composed of $N = 1{,}825$ nodes, as for the yeast PIN. The agreement is very good, considering the relatively large statistical fluctuations we have for this network size. Error bars on the DD model refer to statistical fluctuations on single realizations. It is worth noticing that despite the evident multifractal nature

irregular fractured surfaces, coastlines and other objects that have details over a broad range of scales, and exhibit some kind of self-similarity. Fractals can also be considered as geometrical objects of non-integer dimension. The mass of a steel cube of side L grows in proportion to L^3, reflecting the fact that the cube is a three-dimensional object; in contrast, the mass of a fractal of linear size L grows as L^D, D being the fractal dimension.

A scale-free network can also be considered as a fractal, and its character can be explored through the connectivity distribution. For their network model, Vázquez et al. have analyzed the behavior of the average connectivity $\langle k \rangle$ and of higher moments $\langle k^n \rangle$ as N becomes very large, finding that each of these approaches a finite value for some values of q, but grows without bound for others. For these values, they report the analytical result $\langle k^n \rangle \sim N^{\sigma_n(q)}$, with the exponent $\sigma_n(q)$ given in their equation 2 (simulations confirm this result). This result is significant. Mathematically, if the network were a pure fractal, described by a single fractal dimension, then $\sigma_n(q)$ would depend linearly on n. The nonlinear dependence on n in equation 2 implies that a single fractal dimension does not suffice – the duplication divergence model generates more complex networks having so-called 'multi-fractal' properties. Such networks can be thought of as a statistical mixture of many fractals of different dimensions.

Vázquez et al. have investigated their model further, showing that it also produces other network features that compare favorably to the yeast protein network. This network reveals a number of proteins linked in cyclic fashion into triangles or squares. The duplication-divergence model generates such cycles readily, and the authors compare the networks it generates to the real protein data by considering 'clustering coefficients'. In the case of triangular cycles, one can consider the ratio of the number of triangles to the number of protein triplets that are connected by only two links. This ratio offers a rough measure of the tendency for triangle formation, and a similar approach can be used for squares or higher cycles. With $p = 0.1$ and $q = 0.7$,

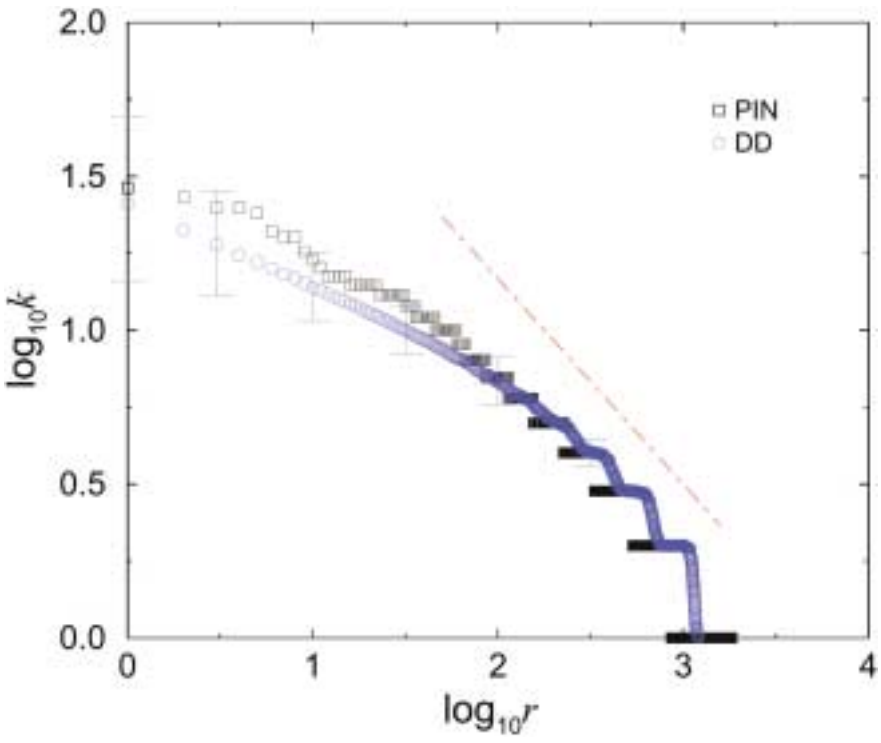

Fig. 2. Zipf plot for the PIN and the DD model with $p = 0.1$, $q = 0.7$ with $N = 1{,}825$. k is the connectivity of a node and r is its rank in decreasing order of k. Error bars represent standard deviation on a single network realization. The straight line is a power law with exponent $1/(1-\gamma)$, with $\gamma = 2.5$, which will correspond to a power-law connectivity distribution $p(k) \sim k^{-\gamma}$.

of the DD model, for a single realization of size consistent with that of the PIN, the intermediate k behavior can be approximated by an effective algebraic decay with exponent $1/(\gamma - 1)$ with $\gamma \simeq 2.5$, as found by Jeong et al. [3]. However, the plot in figure 2 shows a curvature that deviates from the algebraic behavior, evidencing the multifractal nature of the connectivity distribution.

Finally, we examined the bahavior of the DD model under random and selective deletion of nodes and compared it with those obtained for the yeast PIN [3]. Resilience to damage is indeed considered an extremely relevant property of a network. From an applicative point of view it gives a measure of how robust a network is against disruptive modifications, and how far one can go in altering it without destroying its connectivity and therefore functionality. In the random deletion process of a fraction f of nodes and the relevant links, we observed that the network fragments into several disconnected components, with the largest connected component a size of $N(f)$. For random graphs it is known that the fraction of nodes $P(f) = N(f)/N$ belonging to the largest remaining network undergoes an inverse percolation transition [27–29]. In the thermodynamic limit ($N \rightarrow \infty$), above a fraction f_c of deleted nodes, the density P drops to zero; i.e. no dominant network (giant component) is left. On the contrary, in SF networks the density of the largest cluster drops to zero only in the limit $f_c \rightarrow 1$, denoting a high resilience to random damages. Associated with this property, we observed that SF networks are very fragile with respect to targeted removal of the highest connected nodes. In this case a small fraction of removed site fragments completely destroyed the network ($P \rightarrow 0$). A similar behavior has been observed also in the yeast PIN [3]. Figure 3 shows the density of the largest remaining network versus the fraction of removed nodes both for the PIN and the DD model, and for random and selective nodes removal. The latter case consists in systematically removing nodes with the highest degree. The DD network tolerance to damage is determined by the SF nature of its multifractal distribution, and the obtained curves are in very good agreement with the corresponding ones for the yeast PIN. It is worth noting again that the parameters used for the DD network have not been independently estimated, but are those obtained from the previous optimization of the clustering coefficients. The striking analogies in the tolerance behavior are an

these clustering coefficients match well with the data, whereas earlier network models based on preferential attachment produce clustering coefficients about ten times too small for networks of size $N = 1{,}825$.

As one final test, Vázquez et al. study the topological resilience of the networks it generates to the removal of nodes. As they note, this is an important property since it reflects on the network's ability to function in the face of accidental damage or attack. A noteworthy characteristic of scale-free networks is their ability to 'hang together' when sites are removed at random. Even when a large fraction of the nodes have been removed, the network will remain as one more or less fully connected whole. Conversely, scale-free networks are highly vulnerable to intelligent attack, i.e. to a targeted removal of nodes beginning with the most highly connected. Vázquez et al. show that networks created through the duplication-divergence process also work in this fashion, falling apart gracefully in the face of random damage, and collapsing quickly under directed attack (fig. 3). Earlier studies have revealed closely similar behavior in the protein network in yeast.

The results reported here represent an impressive step forward, as they tie basic algorithms of network growth to a real biological basis, naturally reproducing the scale-free network architecture observed in protein networks. At the same time, these results also reflect back on the emerging science of complex networks more generally, as they point the way to networks having richer multi-fractal structures. Complex networks in other settings might well reveal similar characteristics.

Mark Buchanan

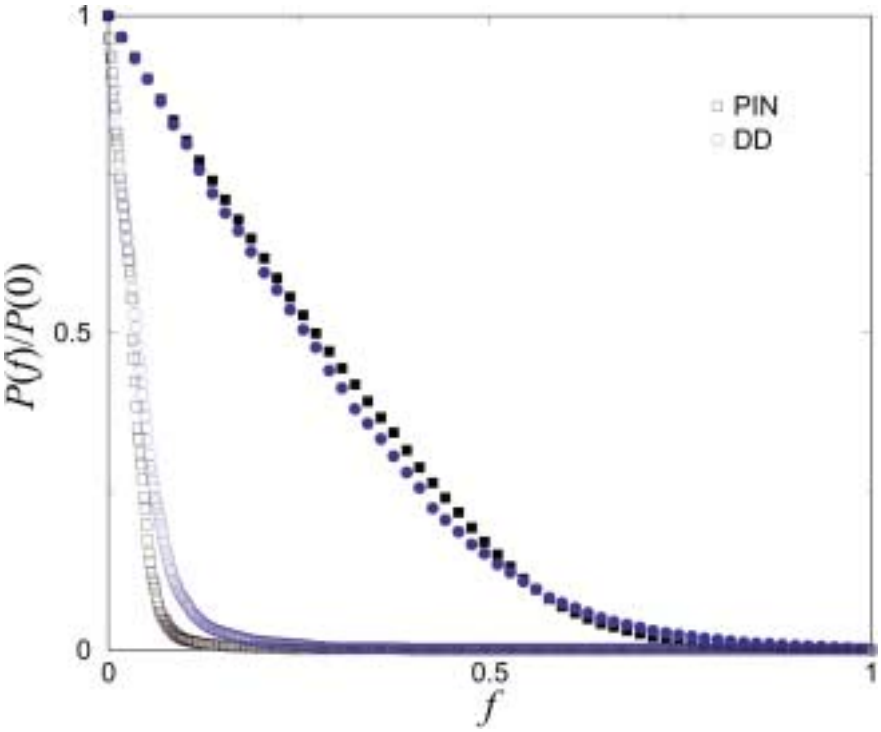

Fig. 3. Fraction of nodes $P(f) = N(f)/N$ in the largest network after a fraction f of the nodes has been removed for the PIN and the DD model. $N(f)$ is defined as the size of the largest network of connected sites. Two different removal strategies have been used, random (filled symbols) and selective (open symbols). In both cases, the DD model curves were obtained after an average of over 100 network realizations.

important test to assess the efficacy of the DD model in reproducing the PIN topology.

In conclusion, we presented a physically motivated dynamic model for the network of protein interactions in biological systems. The model is based on a simple process of gene duplication and differentiation, which is believed to be the main mechanism beyond the evolution of PINs. Although the resulting networks share common features with other SF networks, they present novel and intriguing properties, both in the degree distribution (multifractality) and in their topology. The model reproduces with noticeable accuracy the topological properties of the real PIN of the yeast *S. cerevisiae*.

After submission we became aware of a work treating a model similar to the one introduced here [30].

Acknowledgments

A.M. and A.F. acknowledge funding from Murst Gfin'99. A.V. has been partially supported by the European Network Contract No. ERBFMRXCT980183.

References

1 Uetz PL, et al: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403:623.
2 Wagner A: The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 2001;18:1283–1292.
3 Jeong H, Mason SP, Barabási A-L, Oltvai ZN: Lethality and centrality in protein networks. Nature 2001; 411:41.
4 Watts DJ, Strogatz SH: Collective dynamics of 'small-world' networks. Nature 1998;393:440.
5 Barabási A-L, Albert R: Emergence of scaling in random networks. Science 1999;286:509.
6 Strogatz SH: Exploring complex networks. Nature 2001;410:268.
7 Albert R, Barabási A-L: Topology of evolving networks: Local events and universality. Phys Rev Lett 2000;85:5234.
8 Krapivsky PL, Redner S, Leyvraz F: Connectivity of growing random networks. Phys Rev Lett 2000;85: 4629.
9 Dorogovtsev SN, Mendes JFF: Scaling behaviour of developing and decaying networks. Europhys Lett 2000;52:33.
10 Force A, Lynch M, Pickett, FB, Amores A, Yan Y-I, Postlethwait J: The preservation of duplicate genes by complementary degenerative mutations. Genetics 1999;151:1531.
11 Lynch M, Force A: The probability of duplicate-gene preservation by subfunctionalization. Genetics 2000; 154:459.
12 http://www.biotech.nature.com/web_extras.
13 Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L: Gene families: The taxonomy of protein paralogs and chimeras. Science 1997;278:609.
14 Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. Science 1997;278:631.
15 Ohno S: Evolution by Gene Duplication. Berlin, Springer, 1970.
16 Phizicky EM, Fields S: Protein-protein interactions: Methods for detection and analysis. Microbiol Rev 1995;59:94.
17 von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002;417:399.
18 Yook S-H, Oltvai ZN, Barabási A-L: Functional and topological characterization of protein interaction networks. Preprint.
19 Huynen MA, Bork P: Measuring genome evolution. Proc Natl Acad Sci USA 1998;95:5849.
20 Dorogovtsev SN, Mendes JFF, Samukhin AN: Multifractal properties of growing networks. Europhys Lett 2002;57:334.
21 Albert R, Barabási A-L: Statistical mechanics of complex networks. Rev Mod Phys 2001;74:47.
22 Dorogovtsev SN, Mendes JFF: Evolution of networks. Adv Phys 2002;51:1079.
23 Vázquez A, Flammini A, Maritan A, Vespignani A: Modeling of protein interaction networks. In preparation.
24 Dorogovtsev SN, Mendes JFF: Evolution of networks with aging of sites. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 2000;62:1842.
25 Krapivsky PL, Redner S: Organization of growing random networks. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 2001;63:066123.
26 Newman MEJ, Strogatz SH, Watts DJ: Random graphs with arbitrary degree distributions and their applications. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 2001;64:026118.
27 Albert RA, Jeong H, Barabási A-L: Error and attack tolerance of complex networks. Nature 2000;406:378.
28 Callaway DS, Newman MEJ, Strogatz SH, Watts DJ: Network robustness and fragility: Percolation on random graphs. Phys Rev Lett 2000;85:5468.
29 Cohen R, Erez K, ben-Avraham D, Havlin S: Breakdown of the internet under intentional attack. Phys Rev Lett 2001;86:3682.
30 Solé RV, Pastor-Satorras R, Smith ED, Kepler T: A model of large-scale proteome evolution. Adv Comp Syst 2002;5:43.