Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 352 (2005) 1–27
PHYSICA A
www.elsevier.com/locate/physa

Characterization and modeling of protein–protein interaction networks

Vittoria Colizza[a], Alessandro Flammini[a], Amos Maritan[b], Alessandro Vespignani[a,*]

[a]*School of Informatics and Biocomplexity Center, Indiana University, Bloomington, IN 47408, USA*
[b]*INFM and Dipartment of Physics, Universitá di Padova, Via Marzolo 8, 35131 Padova, Italy*

Available online 13 January 2005

Abstract

The recent availability of high-throughput gene expression and proteomics techniques has created an unprecedented opportunity for a comprehensive study of the structure and dynamics of many biological networks. Global proteomic interaction data, in particular, are synthetically represented as undirected networks exhibiting features far from the random paradigm which has dominated past effort in network theory. This evidence, along with the advances in the theory of complex networks, has triggered an intense research activity aimed at exploiting the evolutionary and biological significance of the resulting network's topology. Here we present a review of the results obtained in the characterization and modeling of the yeast *Saccharomyces Cerevisiae* protein interaction networks obtained with different experimental techniques. We provide a comparative assessment of the topological properties and discuss possible biases in interaction networks obtained with different techniques. We report on dynamical models based on duplication mechanisms that cast the protein interaction networks in the family of dynamically growing complex networks. Finally, we discuss various results and analysis correlating the networks' topology with the biological function of proteins.

PACS: 82.39.Rt; 87.14.Ee; 89.75.Hc

Keywords: Protein interaction networks; Complex networks; Evolution modeling

*Corresponding author.
E-mail address: alessandro.vespignani@th.u-psud.fr (A. Vespignani).

0378-4371/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.physa.2004.12.030

1. Introduction

Complex biological functions in living organisms rarely depend on single components and the possibility of gathering data on the global genomic and proteomic scale has created an unprecedented opportunity to develop comprehensive explanations for biological phenomena. In particular, one of the most important aspects of biological complexity is encapsulated in the structure and dynamics of the many networks emerging at different organizational levels, ranging from intracellular biochemical pathways to ecological interactions [1–6]. While the data sets available to us are often incomplete, yet they suffice for analysis, model development and prediction through model simulations. In addition, the last years have witnessed the developing of a large body of work on the statistical characterization and theory of evolving complex networks, activating an entire research field concerned with the analysis of complex biological networks, in particular focusing on their structure and topology [7–10].

A prominent example in this area is provided by the protein interaction network (PIN) of various organisms which can be mathematically represented as graphs whose nodes symbolize proteins and edges connect pairs of interacting proteins. Global PINs have been collected in particular for the *Saccharomyces cerevisiae* a yeast of the class Hemiascomycetes [11–14], but extensive data are being gathered on higher organisms such as *Drosophila melanogaster* [15]. Noticeably, all interaction data sets exhibit a non-trivial topological structure of the networks, showing a broad connectivity distribution $P(k)$, i.e., the probability that any given protein interacts with k other proteins. This feature implies the statistical abundance of "hubs", that is nodes with a large connectivity, and prompt to a complex architecture that has found further support in the non-trivial correlation and hierarchical features observed in the networks topology [16–18]. Interestingly, these properties are shared by many biological networks that appear to have recurrent architectural principles that might point to common organizational mechanisms [9,10,19]. The resulting networks topology is clearly interwoven with the biological significance of the network's topology and analysis in this direction have indeed pointed out correlation signatures between gene knock-out lethality and the connectivity of the encoded protein [20], negative correlation between the evolution rate of a protein and its connectivity [21,22], and functional constraints in protein complexes [23]. At the same time, topological information is being exploited in predictive methods for protein functional assignment and theoretical models are being developed for the formation of PINs. Despite a careful scrutiny of the possible biases in interaction networks obtained with different techniques is needed [24–26], the results obtained so far on protein interaction networks might open new paths in our understanding of the biological complexity at the "omic" level.

Our aim, here, is to provide an overview of the main results obtained in this area, by privileging the perspective emerged with the use of statistical physics methods and the theory of evolving networks. We start by reviewing the most important experimental techniques used to gather data on protein interactions and the pros and cons as well as the biases intrinsically present in each of them. In the following we

focus on the analysis of the *S. cerevisiae* PIN, presenting a discussion of the topological properties of graphs obtained from different data sets. We then report on works concerning the development of biological evolutionary models that reproduce the structure we observe in PINs. This corresponds to the solution of an "inverse" problem: i.e., given the topology of the network what is the dynamical evolution that gives raise to the observed architectures? This is a fascinating issue since the development of successful models amounts to the understanding of the evolutionary process that has generated the biological life. Finally, we will discuss the approaches developed to facilitate the functional annotation of proteins for which we have few or no functional information at all. In particular we will describe the basic strategy at the basis of global optimization methods that takes into account the whole set of interactions of each uncharacterized protein taking advantage of the information encoded in the connectivity pattern of the whole PIN.

2. Methods

The recent availability of complete genome sequences has pushed forward consistently the development of new high-throughput techniques aimed at detecting protein–protein interactions on a proteome-wide scale. In this section we describe the current state of interaction–detection methods along with a discussion of their positive and negative features. In particular we report on experimental techniques designed to identify physical bindings between proteins (such as yeast two-hybrid systems [11,12] and mass spectrometry analysis of purified complexes of proteins [13,14]), interaction prediction methods whose purpose is to detect functional associations between proteins [27,28], correlated mRNA expression profiles [29,30], genetic interaction–detection [31,32] and in silico approaches (such as gene fusion [33,34], gene neighborhood [35,27] and phylogenetic profiles [36,37]).

2.1. The two-hybrid technique

The two-hybrid technique allows the detection pair-wise protein interactions. It exploits the modular property typical of many eukaryotic transcription factors, which can be usually decomposed in two distinct modules, one directly binding to DNA (DB, DNA-binding domain) and the other activating transcription (AD, transcriptional activating domain) (Fig. 1). The first component, DB, is able to bind to DNA even by itself, while the second module, AD, will activate transcription only if physically associated to a binding domain. This property is the result of a series of analysis made in the 1980's by Ma and Ptashne [38] on transcription factors, while its use for the detection of protein interactions was first proposed in Ref. [39].

As it is illustrated in Fig. 1, in the two-hybrid experiment the test proteins are expressed as fusion proteins (*hybrids*) with a DNA-binding domain (DB, the bait) and a transcriptional activating domain (AD, the prey). Fusions partners are co-expressed in yeast nucleus where a protein–protein interaction is identified thanks to the activation of the reporter gene, which can be detected and measured.

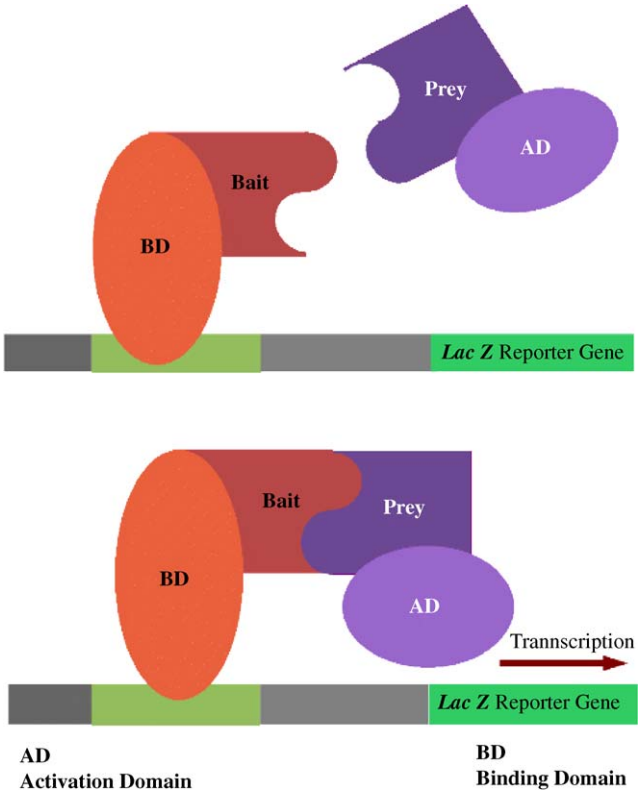

Fig. 1. The two proteins whose interaction is under scrutiny, here indicated as bait and prey, are expressed as fusion proteins, respectively, with a binding domain (BD) and an activation domain (AD). If an interaction between bait and prey takes place, the complex formed activates the transcription of the reporter gene, allowing, as a consequence, the detection of the interaction itself.

The two-hybrid system is able to identify virtually every protein–protein interaction. It is an ex vivo technique that is relatively simple, rapid, and inexpensive, because of the minimal requirements of a two-hybrid screen respect to, e.g., high quantities of purified proteins needed in traditional biochemical approaches. Indeed, it does not require any previous knowledge of the proteins to be tested and can be performed once the corresponding genes are known, thus being suitable for large-scale applications. On the other hand, it only detects binary interactions and does not identify cooperative binding. In addition, some kinds of proteins, such as transcription factors, cannot be studied with this technique since their hybrids could activate the transcription even in absence of any interaction. Furthermore, the extensive use of artificially made hybrids could result in potential drawbacks, by leading to conformational changes in the bait and prey proteins thus preventing transcriptional activation. This is one of the possible causes of false negative interactions, i.e., a true protein–protein interaction which is not detected by

two-hybrid assays. Also, this experimental technique may produce false positives. Indeed, even if two proteins potentially interact into the nucleus, where this techniques takes place, it could happen that they never find close to each other because they could be localized in different cell types or could be expressed in different times of the cell cycle. For this reasons, interactions detected by two-hybrid assays must be critically analyzed in order to assess their biological relevance.

2.2. Protein complex analysis

After the development of ultra-sensitive mass spectrometric techniques for protein identification, new experimental procedures, besides two-hybrid screens, have been used to produce large-scale results for protein–protein interactions, such as purification of protein complexes. This procedure is made up of three main steps: isolation of the bait or target protein, affinity purification of the complex and identification by mass spectrometry of proteins belonging to the complex. The protein of interest is isolated and fused to an affinity tag, by using one of the two protocols: tandem affinity purification (TAP) [13,40] or high-throughput mass-spectrometric protein complex identification (HMS-PCI) [14]. TAP consists of two successive affinity purifications, using two tags fused with the bait and leading to the isolation of the target protein together with its associated proteins. Unfortunately, comparison of results obtained through complex purification with yeast two-hybrid data shows a very small overlap [13]. A possible explanation could rely on the fact that cooperative binding embodied by complexes is not only the result of a sum of pair-wise interactions. Indeed, the main difference between complex purification methods and two-hybrid system relies in the identification of whole complexes isolated in a single step, thus detecting cooperative interactions between proteins which cannot result from two-hybrid screens, where the strategy adopted is based on the bi-modular properties of transcription factors. Moreover, it is an in vivo technique which employs only one artificially made protein (the bait), instead of two as in two-hybrid procedure, thus minimizing possible changes in conformational properties which could lead to steric interference. Complexes are found in physiological settings, since interactions take place in native environment. In order to test the validity of a complex identification, several components of the same complex can be used as tagged baits.

2.3. Interaction detection methods

Besides the physical interactions detected by the high-throughput experimental techniques described above, a complementary insight about protein–protein interactions is given by interaction prediction methods based on genomic information. From the analysis of genome sequences, these methods are able to identify functional associations between proteins (for a review, see Ref. [41]).

Phylogenetic profiles. This approach is based on the simultaneous presence or absence of two proteins in the genomes of different organisms. Functionally interacting proteins indeed tend to have similar phylogenetic profiles [36,37,42].

However, it requires complete sequencing of entire genomes, in order to test the presence or absence of the genes, and is not suitable for essential genes.

Gene fusion. It predicts an interaction between two proteins of a given organism which seem unrelated if they are part of the same polypeptide chain in another organism [33,34].

Gene neighborhood. The conservation of gene neighborhood in the genomes of different organisms is interpreted as an indication of functional association between the proteins encoded by the two genes [27,35]. Indeed, functionally related proteins are often encoded in clusters and this relation is even strengthen by the conservation of such adjacency in different species [43]. However, this happens only in prokaryotic genomes, representing one of the main drawbacks of this approach.

Correlated mRNA expression. This approach predict functional associations between proteins encoded by genes which show similar transcriptional responses to a change in the cellular status [29,30]. Messenger RNA expression profiles can be measured under very different cellular conditions, thus representing an advantage respect to other techniques which can only take into account few settings.

Genetic interaction–detection. Functionally interacting proteins can be detected by synthetic genetic interactions [31,32]. Two non-essential genes show a synthetic lethal interaction if they cause cell death when simultaneously mutated [44,45].

2.4. Data sets comparison and completeness

Several databases have been recently compiled in order to collect and document the vast amount of large-scale data on protein interactions produced in the last few years by high-throughput methods (for a review, see e.g., Ref. [46]). The final aim is a comprehensive characterization of the whole network of connections between proteins resulting by the union of the information gathered in different experiments. On the other hand, data sets comparison and merging must take into account the different conditions under which interactions are detected. Indeed, the intersection between different interaction data shows a surprisingly small overlap [47], underlining the need for a critical evaluation of the biological relevance of large-scale data sets.

The many discrepancies arising from data sets comparison could be due primarily to specific features of experimental methods, each characterized by its own advantages and drawbacks. Results from one method may not largely overlap with those obtained with another technique because of specific restriction and different requirements. In this sense, different experimental techniques could be complementary, thus increasing our knowledge about the network. Secondly, these observations could be the result of low coverage of data sets; i.e., a still partial and incomplete knowledge of the interactions network. Finally, it is also known that results produced by high-throughput techniques, although extensive, may contain spurious interactions (false positives) as well as miss many true interactions (false negatives). The sum of this various factors leads to high uncertainties on data reliability. For instance, even referring to the same experimental technique, yeast two-hybrid assay, one can notice the incredibly small overlap among different data sets [48]; e.g. Ito's

data and Uetz's data share only a very small percentage of interactions, the intersection of the two sets representing, respectively, about 4% and 14% of the total ensembles.

The assessment of the reliability of such data needs a comparison with a trusted reference set, in order to distinguish between validated interactions and background noise. Interactions detected by small-scale experiments could act as a benchmark, since they usually have been thoroughly investigated by multiple experiments and several checks. However, small-scale data sets are not suitable to validate the majority of high-throughput data, because of the very limited number of high-confidence interactions they contain. The same problem is encountered when considering the intersection of different large-scale data sets of protein interactions. Indeed, it has been shown that connections detected by more than one method increase their accuracy with respect to others, while however decreasing their coverage [24], resulting in a very small reference set. For these reasons, the problem of investigating biological relevance and accuracy of protein interactions still represents a crucial step in analyzing protein–protein interaction data.

3. Topological characterization of protein interaction networks

Protein–protein interaction data find an appropriate mathematical representation as undirected graphs whose nodes represent proteins and edges the presence of a direct interaction among them. The statistical analysis of the topology of the resulting graph is therefore a starting point for expressing in concise mathematical terms the hidden regularities and hierarchies of PINs. On their turn, the identification of these features provides information about the organizational principle at the basis of the complicate structure of these graphs, a key point in the connection between the fabric and the biological evolution and function of protein networks.

In view of the discussion concerning the possible biases induced by different experimental techniques, in the following we review the topological properties of three distinct PINs of the yeast *S. cerevisiae* obtained from different data sets:

Network (I): a collection of binary interactions detected by two different two-hybrid assays [11,12], composed of a total of 2831 links among 2152 proteins;

Network (II): interactions obtained from protein complex detection with TAP techniques [13]; it consists of 3221 interactions involving 1361 proteins;

Network (III): a mixed collection of interactions obtained with different experimental techniques, documented at the Database of Interacting Proteins (DIP) [49]; it is composed of 4713 proteins and 14846 interactions. The content of this database is continuously increasing; the number we give here refers at the time we first analyze the data.

It is worth noticing that, while (I) is composed of binary interactions between proteins directly detected by two-hybrid techniques, network (II) assigns hypothetical connections between proteins belonging to the same complex. Indeed, the topology inside a protein complex is not revealed by purification processes: not all

associated proteins will in general interact with the bait, since the interaction could be mediated by other molecules, or interact with the bait at the same time, since interactions could occur under different physiological conditions. Therefore, for a direct comparison with pairwise interactions detected by other experiments, protein complex data have been assigned hypothetical interactions following two different models [47]: the *spoke* model, in which only interactions between the bait and associated proteins occur, and the *matrix* model, which assigns to a given all possible interactions between the proteins belonging to a complex, thus leading to cliques (i.e., fully connected sub-networks). In network (II) we have adopted the spoke model, since it displays a higher accuracy when compared to a reference set [24].

We start by reviewing the properties of the three representative graphs with the analysis of the most basic set of standard metrics. In Table 1 we report the size and the number of interactions of each network, together with the size of the largest (sometime referred to as "giant") component, i.e., the largest connected sub-graph. A first important feature of the graph is highlighted by the average degree, where the degree of a given node is defined as the number of its connections. The average degree is therefore simply $\langle k\rangle = 2l/n$ with l being the total number of links (edges) in the graph and the factor 2 takes into account that each link contributes to the degree of two nodes. The small values of the average degree $\langle k\rangle$, compared to network sizes, states that PINs are sparse graphs. The values observed, however, differ considerably in the three graphs considered, yielding indication that different sampling of the original network are achieved in each data sets.

A more detailed inspection of the graph local cohesiveness is provided by the clustering coefficient. The clustering coefficient measures the local group cohesiveness and is defined for any vertex i as the fraction of connected neighbors of i [50]. Considering a protein i, its clustering C_i is therefore defined as

$$C_i = \frac{2e_i}{k_i(k_i-1)}\ , \tag{1}$$

where e_i is the number of links connecting neighbors of i and $k_i(k_i-1)/2$ is the total number of possible connections among neighbors (for peripheral proteins having $k_i = 1$, C_i is taken equal to zero). A more global quantity for characterizing the graph is the mean clustering coefficient $\langle C\rangle = \frac{1}{n}\sum_i C_i$, where the average is over all the n proteins in the network. This quantity expresses the statistical level of local

Table 1
Average global properties of networks (I), (II) and (III)

	(I)	(II)	(III)
# proteins	2152	1361	4713
# proteins *giant component*	1679 (78%)	1246 (91%)	4626 (98%)
# links	2831	3221	14846
$\langle k\rangle$	2.63	4.73	6.30
$\langle C\rangle$	0.10	0.22	0.09
$\langle C_{rand}\rangle$	0.0064	0.019	0.018

cohesiveness of the graph and it is interesting to compare the empirical measured values with those obtained for random graphs with the same connectivity properties. We compare $\langle C \rangle$ computed on each network with the corresponding average clustering coefficient of a random network with the same degree distribution [51]. In the case of random graphs the clustering coefficient can be expressed in terms of the first and second moment of the distribution [51]:

$$\langle C_{rand} \rangle = \frac{1}{n} \frac{(\langle k^2 \rangle - \langle k \rangle)^2}{\langle k \rangle^3} \tag{2}$$

obtaining the expected values for random graphs with the same properties of the considered data sets. In Table 1 we report the values obtained empirically, and in all cases the measured clustering coefficient is from five to fifteen times larger than the corresponding random one. It is also worth noticing that different levels of cohesiveness exist in the three data sets. While network (II) has the largest clustering coefficient, the larger ratio $\langle C \rangle / \langle C_{rand} \rangle$ is obtained for network (I), indicating, for his data set, the most noticeable departure from the random case. Such a large cohesiveness is a first signature that protein interaction networks do not fit the standard random graph picture, prompting to the presence of organizational principle shaping their structure.

Further differences from the random paradigm emerge by inspecting the distribution of protein degrees, $P(k)$ (Fig. 2), which represents the probability that a randomly chosen protein has a given degree k. Indeed, the observed degree distributions provide a clear mark for a high level of heterogeneity in the connectivity properties. In all data sets the degree distribution is heavy-tailed with a non-negligible probability of having proteins with degree larger than $\langle k \rangle$.

In particular, following Jeong et al. [20], we fit the observed degree distribution to a power-law with exponential cut off

$$P(k) \simeq (k + k_0)^{-\gamma} \mathrm{e}^{-k/k_c} . \tag{3}$$

Degree distributions of (I) and (III) are in good agreement with such functional form—a best fit of real data yields power-law exponents $\gamma^{(\mathrm{I})} \simeq 2.5$ and $\gamma^{(\mathrm{III})} \simeq 2.5$ (slopes of solid lines in Fig. 2, top and bottom), in good agreement with results of Jeong et al. [20] concerning protein interaction data extracted from Ref. [11], and cut offs $k_c^{(\mathrm{I})} \simeq 30$ and $k_c^{(\mathrm{III})} \simeq 100$. Interaction data derived from TAP experiments display a degree distribution which seems to deviate from the behavior observed in (I) and (III), showing the presence of a "bump" in the distribution for intermediate values of the degree. The solid line in Fig. 2 (center) has a slope $\gamma^{(\mathrm{II})} \simeq 2.1$, representing the best fit to the data, using Eq. (3). The high level of heterogeneity embodied by the degree distribution of PINs can be considered as evidence for the absence of a typical scale for the system. In other words, the average degree value $\langle k \rangle$ is not anymore a typical value as in classical random graphs [52,53]. Indeed, the degree fluctuations $\langle k^2 \rangle$ are much larger than the average value, prompting for the presence of overwhelming statistical fluctuations that render the system an example of *scale-free* behavior. It is also worth noticing that the presence of the exponential truncation of the power-law behavior should not be considered in contradiction with

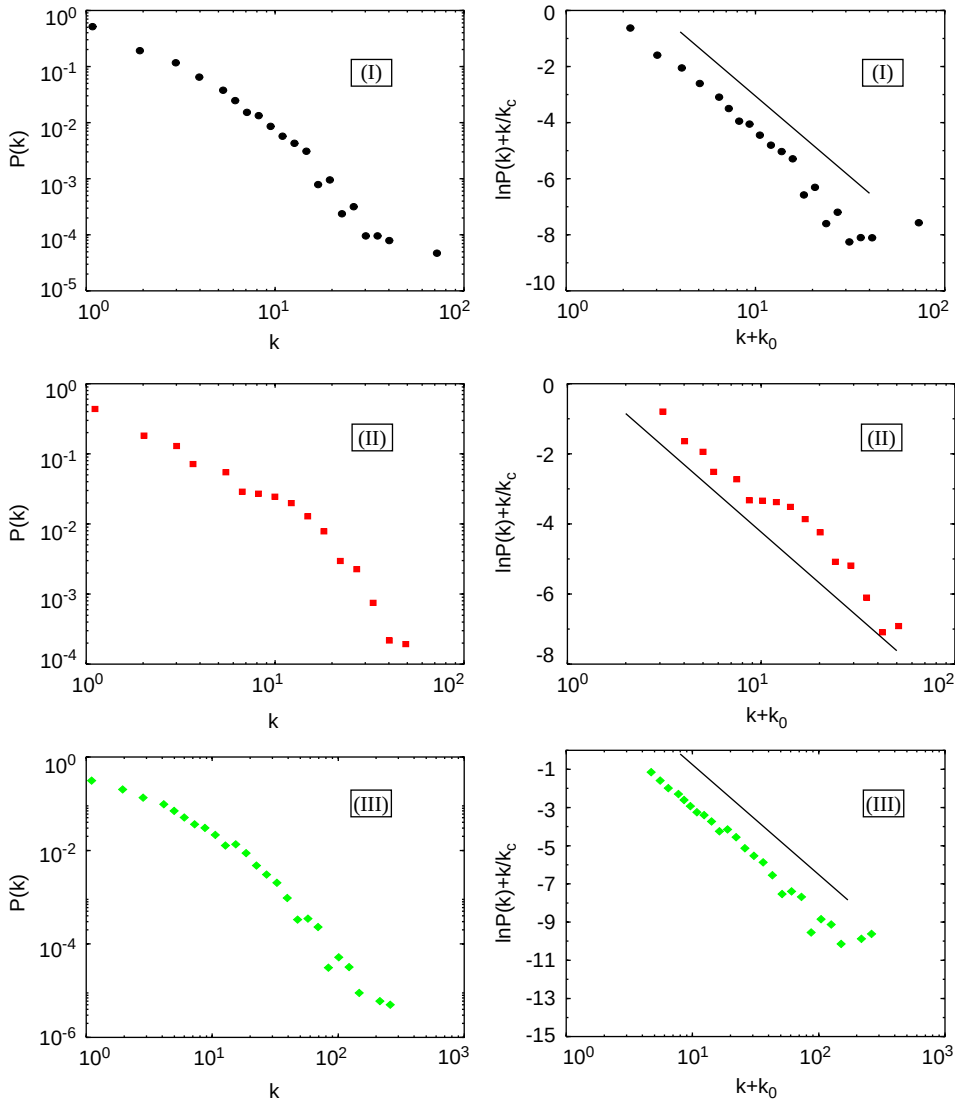

Fig. 2. Degree distribution $P(k)$. On the left we report $P(k)$ in a double log-scale. On the right we plot $\ln P(k) + k/k_c$ (see text) as a function of $k + k_0$ on a single log-scale. From top to bottom: networks (I), (II) and (III). Lines plotted have slopes 2.5, 2.1, 2.5, respectively.

the previous statement. Actually the heavy-tail truncation is the natural effect of the upper limit of the distribution that must necessarily be present in every real-world system.

Along with the vertices hierarchy imposed by the degree distribution, the studied graphs show an architecture imposed by the structural and functional constraints acting on the PINs. In order to uncover this architecture some topological quantities

are customarily studied. A first signature of a hierarchical organization of the network structure can be characterized quantitatively by the clustering coefficient spectrum. This quantity is obtained by averaging the clustering coefficient over vertices with degree k

$$C(k) = \frac{1}{n_k} \sum_i C_i \delta_{k_i,k} \,, \tag{4}$$

where n_k is the number of proteins with degree k. A non-trivial behavior of $C(k)$ provides some hints on the presence of a hierarchy of nodes in the network. In particular, a decaying function $C(k)$ signals a hierarchy in which low-degree proteins belong generally to well interconnected communities (high clustering coefficient) while hubs connect many proteins that are not directly interacting (small clustering coefficient) [17,54–60]. It is natural that the presence of a hierarchy of interconnected proteins group might provide hints on the functional modularity present in the PIN. In Fig. 3 (left) we report results for the clustering spectrum $C(k)$ in the different data sets. Network (II) exhibits a clear heavy-tail which can be fitted to a power-law, $\sim k^{-0.48}$, while networks (I) and (III) do not display a scale-free behavior. Two-hybrid data seem to remain almost constant for small degrees, exhibiting a drop for larger values of k, possibly due to small network size and poor statistics. Finally, the behavior displayed by $C(k)$ for the DIP data set suggests the presence of a structural organization varying continuously over two orders of magnitudes, although do not show a clear functional form for the spectrum $C(k)$. Results observed provide a strong and clear evidence for an inherent hierarchical organization only for network (II), but suggest the presence of a structural organization for the other networks, although characterized by weak and non-univocal signatures.

Another important source of information about the network structural organization lies in the correlations of the connectivities of neighboring proteins. Correlations can be probed by inspecting the average degree of nearest neighbor of a vertex i

$$k_{nn,i} = \frac{1}{k_i} \sum_{j \in nn(i)} k_j \,, \tag{5}$$

where the sum runs on the nearest-neighbors vertices of each vertex i. From this quantity a convenient measure to investigate the behavior of the degree correlation function is obtained by the average degree of the nearest-neighbors spectrum, $k_{nn}(k)$, for vertices of degree k [54,61]

$$k_{nn}(k) = \frac{1}{N_k} \sum_{i|k_i=k} k_{nn,i} \,. \tag{6}$$

This last quantity is related to the correlations between the degree of connected vertices since it can be expressed as

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) \,, \tag{7}$$

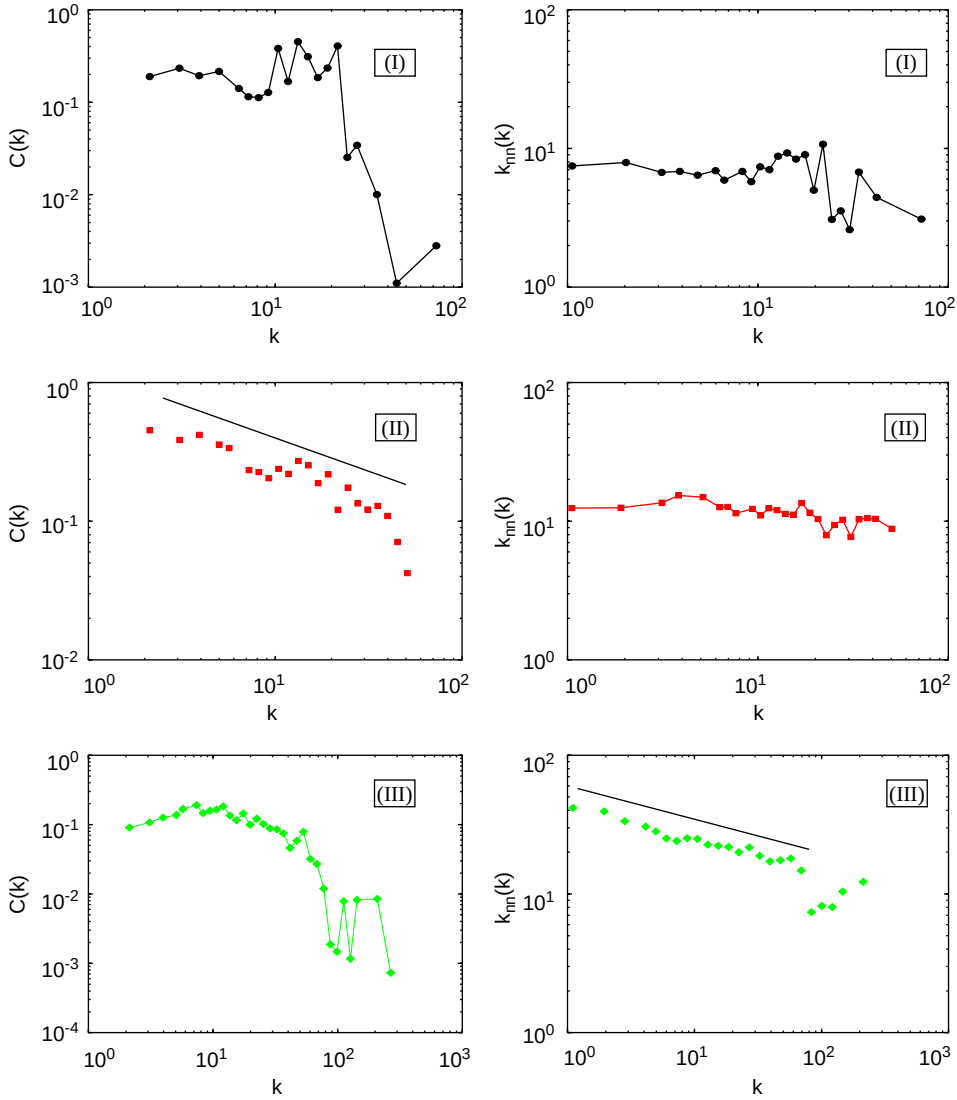

Fig. 3. Average clustering coefficient $C(k)$ (left) and average neighbors degree $k_{nn}(k)$ (right) as a function of protein degree. From top to bottom: networks (I), (II) and (III). Clear power-law behaviors are observed for $C(k)$ in (II), with exponent $\simeq 0.48$, and for $k_{nn}(k)$ in (III), with exponent $\simeq 0.24$.

where $P(k'|k)$ is the conditional probability that any given edge of a protein with degrees k is pointing to a protein with degree k' [18]. If degrees of neighboring vertices are uncorrelated, $P(k'|k)$ is only a function of k' and thus $k_{nn}(k)$ is a constant. When correlations are present, two main classes of possible correlations

have been identified: *assortative* behavior if $k_{nn}(k)$ increases with k, which indicates that large degree vertices are preferentially connected with other large degree vertices, and *disassortative* if $k_{nn}(k)$ decreases with k [62]. In Fig. 3 (right) we plot $k_{nn}(k)$ as a function of protein degree. Evidence of degree correlations are observed only in (III), which exhibits a disassortative behavior with power-law decay with exponent $\simeq 0.24$, whereas (I) and (II) display $k_{nn}(k)$ almost independent of k, thus displaying a lack of correlations. It is worth mentioning that other data sets analysis [18] provide further evidence for disassortative behavior, confirming the presence of non-trivial correlations in the protein interaction connectivity pattern.

While the data sets analysis provides quantitative and even qualitatively different results, it generates convincing evidence that the topology of PINs is departing from the random paradigm entailed by random graph models. Correlations and clustering coefficient prompt to specific organizational principle far beyond the random connectivity pattern. Furthermore, the presence of heavy-tailed distributions represents the potential signature of an emergent behavior leading the network evolution. In other words, the study of the dynamics of the network evolution might shed light on the large-scale structure in which it is organized. These considerations will be basic inputs for the development of evolutionary models as well as relating function and structure in large PINs.

4. Modeling the evolution of the protein interaction network

A satisfactory picture of the basic topological features of PINs described in the previous section cannot be parted by the study of their evolution. Life, and protein networks as a consequence, have not always been as they appear today. It is not unreasonable to think that PINs were originally simpler structures, composed of fewer proteins and possessing a less complex pattern of interactions. In the present section we will show how possibly incorporate in simple networks models the evolutionary forces that have shaped the topology one sees today.

When it was first realized that Yeast's PIN exhibits an algebraic degree distribution and that, therefore, the Random Graph model was unapt to describe it, it did not come completely as a surprise [63]. Several authors had earlier shown that many real networks, noticeably the Internet and WWW, but also others, emerging from the social and the biological sciences, have, to a certain extent, similar properties. The generalized failing of the random graph paradigm became then a loud call for an organizing principle. It had to be of dynamical nature, since all these networks have grown in time by addition of new nodes and links, unsupervised, because of the lack of a "central authority" governing the growth, and, at the same time general enough to embrace the great variety of cases at hand. This void has been somewhat filled by the Barabasi–Albert (BA) model [19], that soon became the paradigm for scale-free networks. Capitalizing on ideas put forward in unsuspected times by Simon [64] and Price [65], Barabasi and Albert proposed a class of growth models for networks based on the principle of

preferential attachment. In practice, a set of $N - m_0$ nodes is sequentially added to an initial core of m_0. When a new node enters the network, carrying along m new edges, it attaches them to m pre-existing nodes, each one chosen randomly with a probability proportional to its degree. This amounts to assume the probability $\Pi_{BA}(k_i) = k_i/\sum_i k_i$ that any given vertex is chosen for the attachment process by the new incoming vertex. The iteration of this simple growth rule leads to networks that asymptotically exhibit a power-law decays as k^{-3}, whose behavior can also computed exactly via a master-equation approach [19,66–68]. Given its simplicity and generality, several generalizations of the model have been developed [68–75] in order to introduce more realistic mechanisms actually occurring in real processes and to extend variability of the power-law exponent. Indeed, the idea of preferential attachment is sound and, infact, general enough to potentially lend its applicability to citation networks, the Internet and the WWW, just to mention few examples. In particular, empirical measurements of degree increase in various experimental data on growing networks support the linear preferential attachment as a realistic approximation of the actual growth dynamics[61,76,77].

While very general, the BA model lacks any biological justification to be claimed as a model for PINs. Furthermore, it cannot reproduce several topological features of PIN beyond the degree distribution. For instance, Samanta and Liang have shown that there is an incredible large number of couple of proteins that share a relevant number of common neighbors, orders of magnitude more than a corresponding, properly fitted, BA or random network. On the other hand, the Barabasi and Albert model is not intended to be a realistic model of any real-world network. Rather it is a zeroth-order conceptual model which can be used as the starting point for models taking into account the various particular processes ruling the dynamics of the network under consideration. In the case of PINs a possible dynamics at the base of the network evolution is found in the classic work of Ohno [78] on the genes duplication and divergence process. It is known that organisms may occasionally pass two copies of one or more genes to their offspring, rarely the entire genome, even though it is known that at least one of such events took place in the history of yeast [79]. The blueprint and the duplicated gene, now distinct, may subsequently follow different evolutionary fate, undergoing to mutations. If the mutation proves to be beneficial the offspring will preserve both genes, now slightly changed. This translates, at the proteome level, in the two genes producing slightly different proteins that will therefore develop a corresponding slightly different pattern of interactions and functionality. On the basis of the duplication divergence mechanism, two simple models of proteome evolution were first developed by Vazquez et al. [80] and by Solé et al. [81] to reproduce topological and large-scaling properties of protein–protein interaction networks and then further developed in Refs. [82,83]. Interestingly enough these models depart consistently from the preferential attachment mechanism. However, as we shall see in the following, they do contain this mechanism as an emergent property and thus providing a basic understanding of its very microscopic origin in the PIN.

Both Vazquez et al. and Solé et al. models are based on the microscopic processes of duplication and complementary degenerative mutations. In Ref. [80] the algorithm for proteome evolution consists of the following rules:

- *Duplication*: a protein i in the network is randomly chosen and duplicated, i.e., a new protein i' is created with links to each neighbor j of the protein i; an interaction between i and i' is created with probability p.
- *Divergence*: each neighbor j is considered; one of its two connections, with i or with i', is chosen and removed with probability q.

The model by Solé et al. [81] takes also into account divergence due to addition of new connections in the network (Fig. 4). The main stream view is that, while the original protein retains its original function (and consequently its pattern of interactions), the duplicated one develops a new, independent and original interaction pattern and function. Recently, however, evidences have been provided [84,85] that, infact, both the duplicated and the parental genes undergo mutations and subfunctionalize: the role played originally by the parental protein is shared between, and possibly optimized by, the two genes together. Both models cited above, in a different measure, spouse this more recent point of view. The process of mutation by addition of new links, anyway, was found to have a probability much smaller than divergence due to deletion [63]. Vazquez et al. have tested the introduction of such mechanism in their model, without obtaining changes in the topological properties. Anyway, in order to have a finite average connectivity in Solé et al. model, the rate of addition of new links α must be inversely proportional to the network size, in agreement with the rates observed in Ref. [63]. While each model uses a slightly different implementation of the basic rules, it is straightforward to see that the biologically motivated local rules actually produce an effective preferential

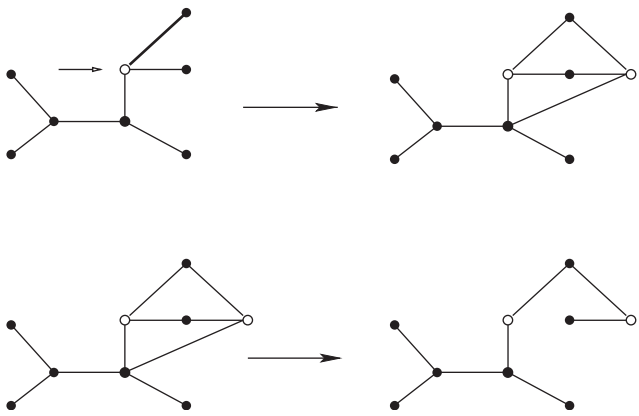

Fig. 4. Duplication–divergence model. Top: a node is duplicated (blank circle, indicated by the arrow). Bottom: some of the original or duplicate edges are removed with probability q.

attachment. Indeed, the probability that a node of the network with degree k gains one more link is given by the probability that one of its neighbors is duplicated (k/N) times the probability of its new link not to be removed ($1-q$). We therefore obtain that, ignoring self-interactions and correlations, each protein has an effective probability of getting new edges given by

$$\Pi(k) \sim (1-q)\,k/N\,, \tag{8}$$

readily disclosing the presence of an emergent preferential attachment in the network dynamics. Indeed, it is interesting to mention that similar copy mechanisms have been implemented also in other domains such as the modeling of the WWW network [86–88], providing a microscopic description of their growth dynamics and the origin of the preferential attachment principle.

PIN models can be analytically studied using a mean-field approximation. Here we follow Ref. [80], since the two approaches lead to analogous results in terms of the respective parameters introduced. The average degree $\langle k\rangle_{N+1}$ of the network with $N+1$ nodes can be expressed as

$$\langle k\rangle_{N+1} = \frac{N\langle k\rangle_N + 2p + (1-2q)\langle k\rangle_N}{N+1}\,, \tag{9}$$

where $2p$ represents the gain in average degree due to self-interacting link and $-2q\langle k_N\rangle$ the loss corresponding to removed connections for divergence process. In the continuum limit for large N, one obtains a differential equation whose solution shows two distinct behaviors, depending on the rate q. For $q>1/2$, a finite average connectivity is reached, i.e., $\langle k\rangle = k_\infty = 2p/(2q-1)$, while for $q<1/2$, $\langle k\rangle$ diverges with N as N^{1-2q}. At $q=q_1=1/2$ a phase transition occurs. Networks obtained display multifractal connectivity properties, with a scale-free behavior characterized by an infinite set of scaling exponents, a features that seems to be related to local inheritance mechanisms [89].

Other relevant quantities have been investigated in Ref. [80], such as the clustering coefficient which displays the correct behavior, reaching a finite value for increasing network size. Once the values of the two rates p and q are set to have clustering coefficient and square coefficient values consistent with those of the protein–protein interaction network of the yeast *S. cerevisiae*, the model is able to reproduce other quantities, such as average degree and degree distribution together with tolerance against random and selective deletion of nodes, which are in good agreement with experimental results (Fig. 5). In Ref. [81], approximate values of the rates δ and β are found by imposing the experimental value of the average degree of the yeast, together with estimations of the ratio α/δ from Ref. [63]. The degree distribution $P(k)$ obtained for networks of size comparable to yeast PINs (Fig. 6) can be fitted by a power-law with an exponential cut off, Eq. (3), already used by Jeong et al. [20] to analyze the connectivity distribution of *S. cerevisiae*. The fit parameters, $\gamma = 2.5 \pm 0.1$ and $k_c \simeq 28$ are in good agreement with those found in Ref. [20]. Other quantities, such as clustering coefficient, average path length and size of the giant component, were quite well reproduced by the model.

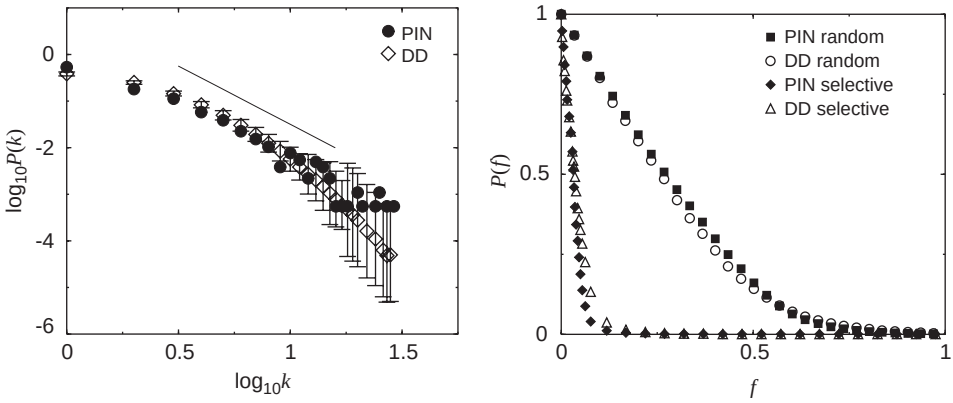

Fig. 5. Duplication–divergence (DD) model results. Left: Connectivity distribution of the protein interaction network (PIN) compared to DD model with optimized rates; error bars represent standard deviations on a single realization. The straight line is a power-law with exponent 2.5. Right: Fraction of nodes $P(f) = N(f)/N$ belonging to the giant component after a fraction f of nodes has been deleted. Comparison of DD model curves (averaged over 100 realizations) with experimental results.

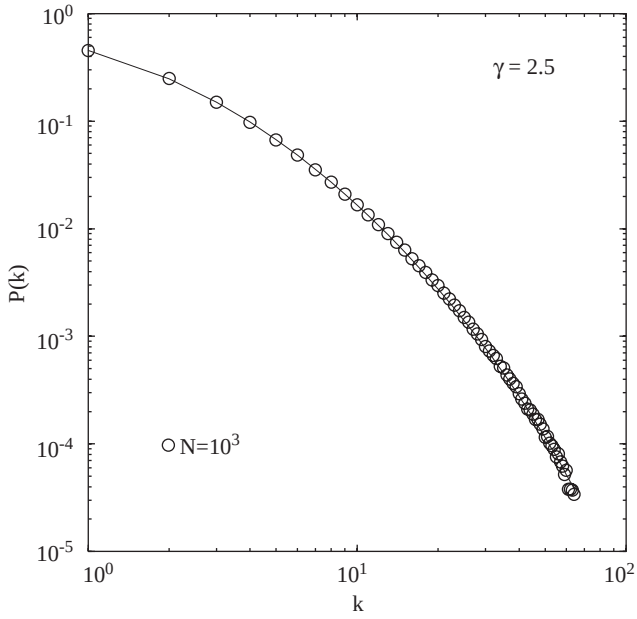

Fig. 6. Degree distribution $P(k)$ for the model [81], averaged over 10^4 realizations of networks with size $N = 10^3$.

Finally, it is worth noticing that it is easy to imagine how the duplication and divergence mechanisms produces a relevant number of proteins that, when compared to a RG or a graph obtained by preferential attachment only, share a large number of common partners, as observed by Samantha and Liang [90]. We will return on the

relevance of this fact when correlation topology–function and methods for function prediction will be discussed in the next section.

5. Proteins functional characterization

With the advent of high-throughput methods, the traditional view of protein function as a task performed by a single protein independently from the others has been substituted by a more general context [91,92], in which interactions between proteins play crucial roles when performing their activities and several cellular processes are the outcome of complex interactions between proteins. The underlying network of interactions thus assumes a deeper meaning in terms of functional relationships between proteins, representing cooperative participation in performing functional tasks.

5.1. Topology/functionality correlations

In the work by von Mering et al. [24] about the quality of different protein interaction data sets in terms of accuracy and coverage, it was shown that in highly accurate data sets functionally related proteins are more likely linked to each other. This feature is usually exploited in function prediction models (see Section 5.2) to infer functional annotation of unclassified proteins from classified neighbors. The authors computed the distribution of interactions according to functional categories and represented the results in terms of a matrix M whose generic element $M(\sigma_i, \sigma_j)$ represents the fraction of links between pairs of proteins performing, respectively, functions σ_i and σ_j. They found that the reference set adopted shows considerably higher values along the matrix diagonal, thus in correspondence of shared functions between proteins.

Here we would like to go further in the investigation about the correlation between the pattern of interactions among proteins and their functionalities, with the purpose of reaching a deepened understanding of biological significance of network architecture. The protein–protein interaction networks of the *S. Cerevisiae* we have considered are those already investigated in Section 3 from a topological point of view, i.e., (I) the two-hybrid data in Refs. [11,12], (II) the data set obtained with an experiment of TAP [13], (III) a heterogeneous collection of interactions detected by different techniques, documented at the Database of Interacting Proteins (DIP) [49]. The functional classification was extracted from the MIPS database [93]; the finest functional classification scheme consists of 424 different functional classes, while the coarse-grained one contains only 18 functional categories. The number of proteins in each data set with no defined functional classification (i.e., belonging to the categories named "*classification not yet clear-cut*" and "*unclassified proteins*") is, respectively: 638 out of 2152 in (I), 279 out of 1361 in (II) and 1665 out of 4713 in (III).

In order to review the likelihood that functionally related proteins are directly connected in the PIN, we compute the rate of interacting protein pairs sharing at

least one functional category, in all three data sets examined. Only classified proteins are taken into account here, together with the whole set of functions they perform. The values obtained adopting the coarse-grained level of functional classification are 83% of interactions between proteins with at least one function in common in two-hybrid data, 83% for TAP and 72% for DIP data set (III). This seems to confirm previous observations, although, the sensitive decrease observed in (III) respect to (I) and (II) highlights the need for caution when interpreting DIP data, since it might indicate the presence of a large amount of false positives.

To determine the actual significance of these results, we compare them with the rates of shared functionalities obtained in two distinct null models, compatible with the constraints embodied by the number of proteins belonging to each functional category. The first null model (NM1) consists in a *functional rewiring* of the network. Starting from the PINs considered, we choose at random two proteins p_i and p_j and exchange their annotations. Unclassified proteins are also considered in the rewiring and the underlying network is not modified. The procedure is repeated a number of times large enough to obtain a network sufficiently "scrambled" but still preserving the composition of each multi-functional annotation. The second null model (NM2), instead, is based on a *random functional assignment* on the empty network. Starting from the network of interactions with no functional annotation, we randomly assign functions to proteins extracted with uniform probability, following three constraints: (a) the number of proteins belonging to each functional category must be globally conserved; (b) a protein cannot be assigned the same function twice; and (c) the number of unclassified proteins must be conserved.

Performing 100 realizations of each null model, we obtain the average values of the rate of interactions between proteins having at least one function in common, together with their standard deviations (see Table 2).

We notice that the random rates of shared functionalities between interacting proteins obtained in the two null models are similar and are both considerably lower than the corresponding real values (exp in Table 2) computed on experimental data. These observations indicate the emergence of a marked correlation between physical link and functional association in protein–protein interaction networks.

The results shown are obtained considering the whole set of classified proteins, independently of their degrees. In order to investigate a possible dependence of the shared functional rate on degree, we have computed the same quantity for low- and

Table 2
Rates of interacting protein pairs sharing at least one functional category. Results obtained from the three networks (exp) are compared with the values averaged over 100 realizations of the two null models, NM1 and NM2, described in the text

$\text{Rate}_{\text{link}\to\text{fcommon}}$	(I) (%)	(II) (%)	(III) (%)
exp	82.90	82.89	72.36
NM1	(60.55 ± 0.19)	(65.35 ± 0.22)	(49.28 ± 0.15)
NM2	(60.64 ± 0.20)	(64.05 ± 0.20)	(49.62 ± 0.16)

Table 3
Comparison among the rates of interacting protein pairs sharing at least one functional category computed on: the whole set of links (link($\forall k, \forall k$)); link(k_{small}, k_{small}) between proteins with *small* degree; link(k_{large}, k_{large}) between proteins with *large* degree; link(k_{small}, k_{large}) between proteins having, respectively, *small* and *large* degree, with the average connectivity $\langle k \rangle$ being the separation value. For comparison, we report also the values obtained with the two null models—NM1 and NM2

$\text{Rate}_{\text{link}\to\text{fcommon}}$	(I) (%)	(II) (%)	(III) (%)
link($\forall k, \forall k$)	82.90	82.89	72.36
link(k_{small}, k_{small})	80.85	81.16	63.71
link(k_{large}, k_{large})	88.50	85.33	78.76
link(k_{small}, k_{large})	75.30	79.70	63.17
NM1	(60.55 ± 0.19)	(65.35 ± 0.22)	(49.28 ± 0.15)
NM2	(60.64 ± 0.20)	(64.05 ± 0.20)	(49.62 ± 0.16)

high-connectivity proteins. The average connectivity $\langle k \rangle$ is the separation value: k_{small} indicates low degrees ($k_{small} < \langle k \rangle$), k_{large} refers to high degrees ($k_{large} > \langle k \rangle$). In Table 3 we report results obtained from the experimental data corresponding to networks (I), (II) and (III) and compare them with results from the null models. No distinctions based on protein connectivity are considered in the null models, since functional annotation is by definition uncorrelated with topology.

A common behavior can be observed in all networks: the rate of functional commonality between interacting proteins increases when considering two proteins with large degree k_{large}, while it is considerably lower when the connected pair is composed at least by a protein with small degree k_{small}. The lowest value is assumed in correspondence of the type of links (k_{small}, k_{large}). We have also investigated the role of peripheral proteins ($k = 1$), being affected by false interactions with higher probability. We have thus computed the same quantities as before without including peripheral proteins among those with k_{small}. The observed trend is unchanged, showing a deeper correlation of functional characterization with topology, which should be investigated in the next future. As we will see in the next section this general feature will be exploited in protein function prediction methods.

5.2. Function prediction methods

Despite the impressive progresses performed during the last years in genome sequencing and high-throughput proteomics techniques, a great amount of encoded proteins per completely sequenced genome is still functionally uncharacterized [94], so that the development of bioinformatics methods for function prediction assumes a crucial importance in order to fully exploit genome data. The list of available in silico approaches to protein function prediction is extensive and includes methods based on sequence similarity, clustering patterns of co-regulated genes [95,96], phylogenetic profiles [97] and analysis of protein complexes [13,14]. In this section we will focus on

those methods [11,12,90,98–101] that more closely rely on the assumed correlation between function and topology for which evidences have been discussed in the previous section.

The basic strategy usually relies on the assumption that two proteins will be more likely functionally related when they are close to each other, rather than if they are far away in the network. This assumption has been first exploited in the "majority rule" method [98,99], whose functional assignment for an uncharacterized protein is obtained by looking at the functional annotations of its classified neighbors. The assignment proposed consists of the most common function(s) among the ones performed by the classified binding partners.

Capitalizing on the majority rule method, Samantha and Liang [90] assumes that couples of proteins that share a large number of common neighbors are most likely to be functionally related and therefore the function of one of the two member of the couple can be deduced if that of the other is known. The mechanism of duplication and divergence discussed earlier provides a rationale to this method: proteins sharing a large number of common partners are also those that more recently diversified from a common ancestor and therefore those whose functions have emerged as a subfunction of the common parent's function. A further modification of the basic majority rule strategy is proposed in Ref. [101] where predictions are made on the basis of a tree-like network of interactions constructed with the introduction of a new definition of "functional" distance between proteins. The latter takes into account not only the shortest path between the proteins but even their number of common partners.

All the methods above offer insights on the structure of protein network functionality from a different point of view and have their own range of applicability. A method that appears to have a more general and automatic applicability is the "global optimization model" (GOM) described in Ref. [100]. Leveraging on the same assumptions on which the "majority rule" [99] is based on, it addresses one of its main weakness, namely the fact that the majority rule completely disregards the role played by the unclassified proteins in the function assignment process. In poorly annotated interactomes, infact, the potential information conveyed by the unknown proteins could easily overweight that from the un-annotated ones. More specifically, the functional assignment GOM gives to an unclassified protein depends on the annotations of the entire set of its binding partners, considering both classified and unclassified proteins. The method is thus able to provide an assignment for all uncharacterized proteins in the network, in a self-consistent way, taking into account also the functional information carried by connections between proteins of unknown function. A score E is associated to each functional assignment, by looking at shared functionalities between connected proteins:

$$E = -\sum_{i<j} J_{ij}\delta_{\sigma_i,\sigma_j} - \sum_i h_i(\sigma_i)\,, \tag{10}$$

where $J_{ij} = 1$ only if the two proteins i and j are both unclassified and directly connected, otherwise $J_{ij} = 0$, σ_i is the function of protein i, $\delta_{l,m}$ is the discrete delta function and $h_i(\sigma_i)$ represents the number of classified binding partners of protein i sharing the same function σ_i. The score assigns value -1 to each connection between

unclassified proteins or between classified and unclassified proteins when sharing a common function. The functional assignment proposed by the method is the one that minimizes the cost function E over the whole network, thus performing a global optimization, while majority rule algorithm, ignoring connections between proteins with unknown functions, is obtained by minimizing the second term only of the right-hand side of Eq. (10). Interestingly, the optimization process leads to several equivalent or nearly equivalent minima corresponding to equal or very close values of the score function. Indeed, the computational problem is frustrated because of the presence of the boundary conditions imposed by the classified proteins, which do not allow to satisfy all requested shared functionalities between interacting proteins. The functional annotations proposed by the GOM are those of the optimal configurations which correspond to the minima of the score function evaluated on the whole network, therefore allowing in a natural way for multiple assignments.

Several quantities have been investigated in order to test the functional predictions proposed by the global optimization method. The predictive accuracy has been evaluated by testing the method on a set of classified proteins and quantified by the introduction of a measure of the rate of successful predictions, which corresponds to the percentage of proposed assignments actually recovering correct annotations of the test proteins. The accuracy of the method depends on several factors, including the classification scheme adopted, the particular network considered and on the fraction of classified proteins present, but is surprisingly reliable. Results show that accuracy increases with the degree of the unclassified protein under consideration and approach the maximum achievable when the coarsest functional scheme and well connected proteins are considered. Also, the method proved very robust against the presence of false positives/negatives links by which data sets are certainly affected. The promising results obtained with GOM have recently motivated further studies in this direction aimed at modifying the score function in Eq. (10) in order to take more throughly into account the topological structure of the network and the observed correlation between interacting proteins [102].

6. Outlook

As most often in the biological context, results concerning the analysis of PINS are not exempt from various caveats. Nevertheless, the global statistical analysis of the ever increasing number of data sets available provides a general conceptual framework in which eventually many questions concerning evolution, biological function and the network topology may be addressed. Indeed, several groups have started to obtain interesting findings concerning the modular architecture of PINs and its role in the evolutionary process. For instance, specific topological motifs of the protein interaction networks might be associated with functional modules providing a first step in connecting the biological networks topological architecture to their detailed function and evolution. At the same time, genome sequencing and PINs data gathering of different organisms open exciting opportunities for comparative analysis and evolutionary studies [6,21–23]. The detection of the

particular role at the evolutionary level of proteins belonging to specific topological motifs finds an immediate impact in evolutionary models for the PIN based on duplication–divergence mechanisms. As well, the relevance of specific motifs and local configuration analysis aimed at identifying cellular function modules may enrich local and global algorithm for protein function assignment based on the PIN. In this perspective, the results reviewed in this paper might represent a relevant contribution toward the answer of specific and detailed questions about biological complexity.

Acknowledgements

We thank A. Vazquez and M.Vergassola for discussion and data sharing. A.V. is partially funded by the EC-FOP COSIN-IST-2001-33555.

References

[1] L.H. Hartwell, et al., From molecular to modular cell biology, Nature 402 (1999) C47.
[2] A. Wagner, D.A. Fell, The small world inside large metabolic networks, Proc. Roy. Soc. London Ser. B 268 (2001) 1803.
[3] Y.I. Wolf, G.P. Karev, E.V. Koonin, Scale-free networks in biology: new insights into the fundamentals of evolution?, Bioessays 24 (2002) 105.
[4] E.V. Koonin, Y.I. Wolf, G.P. Karev, The structure of the protein universe and genome evolution, Nature 420 (2002) 218.
[5] U. Alon, Biological networks: the tinkerer as an engineer, Science 301 (2003) 1866.
[6] A.-L. Barabasi, Z.N. Oltvai, Network biology: understanding the cell's functional organization, Nat. Rev. Genet. 5 (2004) 101.
[7] R. Albert, A.-L. Barabasi, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 47.
[8] M. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167.
[9] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of Networks, Oxford Univesrity Press, Oxford, 2003.
[10] R. Pastor-Satorras, A. Vespignani, Evolution and Structure of the Internet, Cambridge Univesrity Press, Cambridge, 2004.
[11] P. Uetz, et al., A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*, Nature 403 (2000) 623.
[12] T. Ito, et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl. Acad. Sci. USA 98 (2001) 4569.
[13] A.C. Gavin, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (2002) 141.
[14] Y. Ho, et al., Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, Nature 415 (2002) 180.
[15] L. Giot, et al., A protein interaction map of *Drosophila melanogaster*, Science 302 (2003) 1727.
[16] S.Y. Yook, Z.N. Oltvai, A.-L. Barabási, Functional and topological characterization of protein interaction networks, Proteomics 4 (2004) 928.
[17] E. Ravasz, A.-L. Barabási, Hierarchical organization in complex networks, Phys. Rev. E 67 (2003) 026112.
[18] S. Maslov, K. Sneppen, Specificity and stability in topology of protein newtorks, Science 296 (2002) 210.
[19] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509.

[20] H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41.
[21] H.B. Fraser, et al., Evolutionary rate in the protein interaction network, Science 296 (2002) 750.
[22] H.B. Fraser, et al., A simple dependence between protein evolution rate and the number of protein–protein interactions, BMC Evol. Biol. 3 (2003) 11.
[23] S. Wuchty, et al., Evolutionary conservation of motif constituents in the yeast protein interaction network, Nat. Genet. 35 (2003) 176.
[24] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork, Comparative assessment of large-scale data sets of protein–protein interactions, Nature 417 (2002) 399.
[25] J.D. Bloom, C. Adami, Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interaction data sets, BMC Evol. Biol. 3 (2003) 21.
[26] I.K. Jordan, et al., No simple dependence between protein evolution rate and the number of protein–protein interactions, BMC Evol. Biol. 3 (2003) 1.
[27] T. Dandekar, B. Snel, M. Huynen, P. Bork, Conservation of gene order: a fingerprint of proteins that physically interact, Trends Biochem. Sci. 23 (1998) 324.
[28] H. Ge, Z. Liu, G.M. Church, M. Vidal, Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*, Nat. Genet. 29 (2001) 482.
[29] R.J. Cho, et al., A genome-wide transcriptional analysis of the mitotic cell cycle, Molecular Cell. 2 (1998) 65.
[30] T.R. Hughes, et al., Functional discovery via a compendium of expression profiles, Cell 102 (2000) 109.
[31] A.H. Tong, et al., Systematic genetic analysis with ordered arrays of yeast deletion mutants, Science 294 (2001) 2364.
[32] H.W. Mewes, et al., MIPS: a database for genomes and protein sequences, Nucleic Acids Res. 30 (2002) 31.
[33] A.J. Enright, I. Iliopoulos, N.C. Kyrpides, C.A. Ouzounis, Protein interaction maps for complete genomes based on gene fusion events, Nature 402 (1999) 86.
[34] E.M. Marcotte, M. Pellegrini, H.L. Ng, D.W. Rice, D.O. Yeates, D. Eisenberg, Detecting protein function and protein–protein interactions from genome sequences, Science 285 (1999) 751.
[35] R. Overbeek, M. Fonstein, M. D'Souza, G.D. Pusch, N. Maltsev, The use of gene clusters to infer functional coupling, Proc. Natl. Acad. Sci. USA 96 (1999) 2896.
[36] M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, T.O. Yeates, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, Proc. Natl. Acad. Sci. USA 96 (1999) 4285.
[37] M.A. Huynen, P. Bork, Measuring genome evolution, Proc. Natl. Acad. Sci. USA 95 (1998) 5849.
[38] J. Ma, M. Ptashne, A new class of yeast transcriptional activators, Cell 51 (1987) 113.
[39] S. Fields, O.-K. Song, A novel genetic system to detect protein–protein interactions, Nature 340 (1989) 245.
[40] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Seraphin, A generic protein purification method for protein complex characterization and proteome exploration, Nat. Biotechnol. 17 (1999) 1030.
[41] A. Valencia, F. Pazos, Computational methods for the prediction of protein inetarctions, Curr. Opin. Struct. Biol. 12 (2002) 368.
[42] T. Gaasterland, M.A. Ragan, Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes, Microb. Comp. Genomics 3 (1998) 199.
[43] J. Tamames, G. Casari, C. Ouzounis, A. Valencia, Conserved clusters of functionally related genes in two bacterial genomes, J. Mol. Evol. 44 (1997) 66.
[44] P. Novick, B.C. Osmond, D. Botstein, Suppressors of yeast actin mutations, Genetics 121 (1989) 659.
[45] L. Guarente, Synthetic enhancement in gene interaction—a genetic tool come of age, Trends Genet. 9 (1993) 362.
[46] I. Xenarios, D. Eisenberg, Protein interaction databases, Curr. Opin. Biotechnol. 12 (2001) 334.

[47] G.D. Bader, C.W.V. Hogue, Analyzing yeast protein–protein interaction data obtained from different sources, Nat. Biotechnol. 20 (2002) 991.
[48] C.M. Deane, L. Salwinski, I. Xenarios, D. Eisenberg, Protein interactions—two methods for assessment of the reliability of high throughout observations, Mol. Cell. Proteomics 1 (2002) 349.
[49] Database of Interacting Proteins, http://dip.doe-mbi.ucla.edu/.
[50] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[51] M.E.J. Newman, Random graphs as models of networks, in: S. Bornholdt, H.G. Schuster (Eds.), Handbook of Graphs and Networks: from the Genome to the Internet, Wiley-VCH, Berlin, 2003, pp. 35–68.
[52] P. Erdös, A. Rényi, On random graphs, Publicationes Mathematicae 6 (1959) 290.
[53] B. Bollobás, Random Graphs, Cambridge University Press, Cambridge, 2001.
[54] A. Vazquez, R. Pastor-Satorras, A. Vespignani, Large-scale topological and dynamical properties of Internet, Phys. Rev. E 65 (2002) 066130.
[55] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Pseudofractal scale-free web, Phys. Rev. E 65 (2002) 066122.
[56] S. Jung, S. Kim, B. Kahng, Geometric fractal growth model for scale-free networks, Phys. Rev. E 65 (2002) 056101.
[57] G. Caldarelli, R. Pastor-Satorras, A. Vespignani, Structure of cycles and local ordering in complex networks, Eur. Phys. J. B 36 (2003) 203.
[58] G. Bianconi, A. Capocci, Number of loops of size h in growing scale-free networks, Phys. Rev. Lett. 90 (2003) 078701.
[59] E. Ravasz, A.L. Somera, D.A. Mongru, Z. Oltvai, A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (2002) 1551.
[60] A. Vazquez, Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations, Phys. Rev. E 67 (2003) 056104.
[61] R. Pastor-Satorras, A. Vazquez, A. Vespignani, Dynamical and correlation properties of the Internet, Phys. Rev. Lett. 87 (2001) 258701.
[62] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701.
[63] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mol. Biol. Evol. 18 (2001) 1283.
[64] H.A. Simon, On a class of skew distribution functions, Biometrika 42 (1955) 425.
[65] D.J. de, S. Price, A general theory of bibliometric and other cumulative advantage processes, J. Am. Soc. Inform. Sci. 27 (1976) 292.
[66] A.-L. Barabási, R. Albert, H. Jeong, Mean-field theory for scale-free random networks, Physica A 272 (1999) 173.
[67] P.L. Krapivsky, S. Redner, F. Leyvraz, Connectivity of growing random networks, Phys. Rev. Lett. 85 (2000) 4629.
[68] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. 85 (2000) 4633.
[69] P.L. Krapivsky, S. Redner, Organization of growing random networks, Phys. Rev. E 63 (2001) 066123.
[70] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Principles of statistical mechanics of uncorrelated random networks, Nucl. Phys. B 666 (2003) 396.
[71] S.N. Dorogovtsev, J.F.F. Mendes, Scaling behaviour of developing and decaying networks, Europhys. Lett. 52 (2000) 33.
[72] P.L. Krapivsky, S. Redner, A statistical physics perspective on Web growth, Comput. Netw. 39 (2002) 261.
[73] G. Bianconi, A.-L. Barabási, Bose-Einstein condensation in complex networks, Phys. Rev. Lett. 86 (2001) 5632.
[74] G. Bianconi, A.-L. Barabási, Competition and multiscaling in evolving networks, Europhys. Lett. 54 (2001) 436.
[75] G. Ergün, G.J. Rodgers, Growing random networks with fitness, Physica A 303 (2002) 261.

[76] H. Jeong, Z. Néda, A.-L. Barabási, Measuring preferential attachment in evolving networks, Europhys. Lett. 61 (2003) 567.
[77] M.E.J. Newman, Clustering and preferential attachment in growing networks, Phys. Rev. E 64 (2001) 025102.
[78] S. Ohono, Evolution by Gene Duplication, Springer, Berlin, 1970.
[79] K.H. Wolfe, D.C. Shields, Molecular evidence for an ancient duplication of the entire yeast genome, Nature 387 (1997) 708.
[80] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Modeling of protein interaction networks, ComPlexUs 1 (2003) 38.
[81] R.V. Solé, R. Pastor-Satorras, E. Smith, T.B. Kepler, A model of large-scale proteome evolution, Adv. Complex Systems 5 (2002) 43.
[82] A. Bhan, D.J. Galas, T.G. Dewey, A duplication growth model of gene expression networks, J. Mol. Biol. 18 (2002) 1486.
[83] A. Wagner, How the global structure of protein interaction networks evolves, Proc. Roy. Soc. London B 270 (2003) 457.
[84] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y.-l. Yan, J. Postlethwait, Preservation of duplicate genes by complementary, degenerative mutations, Genetics 151 (1999) 1531.
[85] M. Lynch, A. Force, The probability of duplicate gene preservation by subfunctionalization, Genetics 154 (2000) 459.
[86] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A.S. Tomkins, The web as a graph: measurements models and methods, Proceedings of the 32nd Annual ACM Symposium on Theory of Computing, 2000, p. 163.
[87] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A.S. Tomkins, E. Upfal, Stochastic models for the web graph, Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science, 2000, p. 57.
[88] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web, Comput. Netw. 33 (2000) 309.
[89] S.N. Dorogovtsev, A.N. Samukhin, J.F.F. Mendes, Multifractal properties of growing networks, Europhys. Lett. 57 (2002) 334.
[90] M.P. Samanta, S. Liang, Predicting protein functions from redundancies in large-scale protein interaction networks, Proc. Natl. Acad. Sci. USA 100 (2003) 12579.
[91] T.C. Hodgman, A historical perspective on gene/protein functional assignment, Bioinformatics 16 (2000) 10.
[92] D. Eisenberg, E.M. Marcotte, I. Xenarios, T.O. Yeates, Protein function in post-genomic era, Nature 405 (2000) 823.
[93] The MIPS Comprehensive Yeast Genome Database (CYGD), http://mips.gsf.de/proj/yeast/CYGD/db/.
[94] H.W. Mewes, K. Albermann, K. Heumann, S. Liebl, F. Pfeiffer, MIPS: a database for protein sequences, homology data and yeast genome information, Nucleic Acids Res. 25 (1997) 28.
[95] M.Q. Zhang, Promoter analysis of co-regulated genes in the yeast genome, Comput. Chem. 23 (1999) 233–250.
[96] H.C. Harrington, C. Rosenow, J. Relief, Monitoring gene expression using DNA microarrays, Curr. Opin. Microbiol. 3 (2000) 285–291.
[97] M. Pellegrini, E. Marcotte, M.J. Thompson, D. Eisemberg, T.O. Yeates, Assigning protein functions by comparative analysis: protein phylogenetic profile, Proc. Natl. Acad. Sci. USA 96 (1999) 4285–4288.
[98] B. Schwikowski, P. Uetz, S. Fields, A network of protein–protein interactions in yeast, Nat. Biotechnol. 18 (2000) 1257.
[99] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, T. Tagaki, Assessment of prediction accuracy of protein function from protein–protein interaction data, Yeast 18 (2001) 523.
[100] A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, Global protein function prediction from protein–protein interaction network, Nat. Biotechnol. 21 (2003) 697.

[101] C. Brun, F. Chevenet, D. Martin, J. Wojcik, A. Guenoche, B. Jacq, Functional classification of proteins for the prediction of cellular function from a protein–protein interaction network, Genome Biol. 5 (2003) R6.
[102] V. Colizza, P. De Los Rios, A. Flammini, A. Maritan, Protein function prediction from protein–protein interaction data, under review of Genome Biology.