RESEARCH ARTICLE

Virtual identification of essential proteins within the protein interaction network of yeast

Ernesto Estrada

Complex Systems Research Group, X-Ray Unit, RIAIDT, University of Santiago de Compostela, Edificio CACTUS, Santiago de Compostela, Spain

Received: April 7, 2005
Revised: May 23, 2005
Accepted: May 31, 2005

Topological analysis of large scale protein-protein interaction networks (PINs) is important for understanding the organizational and functional principles of individual proteins. The number of interactions that a protein has in a PIN has been observed to be correlated with its indispensability. Essential proteins generally have more interactions than the nonessential ones. We show here that the lethality associated with removal of a protein from the yeast proteome correlates with different centrality measures of the nodes in the PIN, such as the closeness of a protein to many other proteins, or the number of pairs of proteins which need a specific protein as an intermediary in their communications, or the participation of a protein in different protein clusters in the PIN. These measures are significantly better than random selection in identifying essential proteins in a PIN. Centrality measures based on graph spectral properties of the network, in particular the subgraph centrality, show the best performance in identifying essential proteins in the yeast PIN. Subgraph centrality gives important structural information about the role of individual proteins, and permits the selection of possible targets for rational drug discovery through the identification of essential proteins in the PIN.

Keywords:
Centrality measures / Complex networks / Essential proteins / Protein-protein interactions / Spectral graph theory

1 Introduction

A recent explosion of research papers related to the structure of complex networks has led to important results related to the topological properties of biological networks [1]. These networks, which include metabolic networks [2, 3] and protein interaction networks (PINs) [4–6], share important structural features with other real-world networks in disparate fields ranging from the Internet to social networks [7–9]. One of the most important topological characteristics shared by almost all complex networks is the so-called "small-worldness" [10]. Other properties, such as "scale-freeness" [11] and modularity [12, 13] have also been reported for metabolic networks and PINs. On the other hand, the number of links per node (node degree) in the PIN of *Saccharomyces cerevisiae* has been observed to be correlated with the lethality of removing such proteins from the PIN [14]. According to this result those nodes with a large number of links, the so-called "hubs", tend to be essential. The indispensability of a gene defines its functional significance and when such genes are knocked out the cell becomes unviable [15]. The yeast PIN has also been the objective of several other topological analyses aimed at the detection of protein functionality and evolution [16–19].

Node degree is one of the several local topological properties of networks that are known as "centrality measures". The notion of centrality comes from its use in social net-

Correspondence: Ernesto Estrada, Complex Systems Research Group, X-Ray Unit, RIAIDT, University of Santiago de Compostela, Edificio CACTUS, Campus Sur, Santiago de Compostela 15782, Spain
E-mail: estrada66@yahoo.com
Fax: +34-981-547-077

Abbreviations: BC, betweenness centrality; **CC**, closeness centrality; **DC**, degree centrality; **EC**, eigenvector centrality; **IC**, information centrality; **PIN**, protein-protein interaction network; **SC**, subgraph centrality

works [20]. Intuitively, it is related to the ability of a node to communicate directly with other nodes, or to its closeness to many other nodes or to the quantity of pairs of nodes that need a specific node as intermediary in their communications [21]. These ideas have materialized in some well-known centrality measures such as degree centrality (DC), closeness centrality (CC), and betweenness centrality (BC) [20, 22]. Other centrality measures, known as eigenvector centrality (EC) and information centrality (IC), were developed by Bonacich [23, 24] and Stephenson and Zelen [25], respectively. More recently Estrada and Rodríguez-Velázquez [26] introduced a centrality measure that accounts for the weighted participation of nodes in all subgraphs of the network.

Although some centrality measures have been analyzed for biological networks [27, 28], a systematic study of the relationships between centralities and protein indispensability in PINs has not been reported. It has been argued that "the biological consequences of these topological properties are not clear – in fact, there might not be any, as both small-world and scale-free behavior can be explained by well-known evolutionary events" [5]. However, we consider here a more pragmatic approach. The main objective of this work lies in analysis of the efficacy of different approaches based on centrality measures to identify lethal proteins in a protein-protein interaction network. If we have determined the PIN of a pathogenic organism for which we are interested in designing a new drug, our next objective will be to select some proteins from the organism that can be used as targets for this new drug to be designed. In other words, we need to select some essential proteins that can be attacked by the new drug, thus killing the pathogenic organism. To determine which centrality measure is more appropriate for selecting essential proteins in a PIN, we have studied the relationships between DC, CC, BC, EC, IC and subgraph centrality (SC) with essentiality of proteins in the yeast PIN.

2 Materials and methods

2.1 Centrality measures

In the context of PINs, we will refer to "protein centrality" to characterize the importance or contribution of an individual protein to the global structure or configuration of the PIN.

The simplest of all centrality measures is the DC. $DC(i)$ is the number of ties incident upon a node i, *i.e.*, the number of proteins that are interacting with protein i. Another centrality measure is the CC of a connected network. The CC of protein i is the sum of graph-theoretic distances from all other proteins in the PIN, where the distance $d(i, j)$ from one protein i to another j is defined as the number of links in the shortest path from one to the other. The CC of protein i in a PIN is given by the following expression:

$$CC(i) = \frac{N-1}{\sum_{j} d(i,j)} \tag{1}$$

Another popular centrality measure is BC [29]. BC characterizes the degree of influence a protein has in "communicating" between protein pairs and is defined as the fraction of shortest paths going through a given node. If $\rho(i, j)$ is the number of shortest paths from protein i to protein j, and $\rho(i, k, j)$ is the number of these shortest paths that pass through protein k in the PIN, then the BC of node k is given by:

$$BC(k) = \sum_{i} \sum_{j} \frac{\varrho(i,k,j)}{\varrho(i,j)}, i \neq j \neq k \tag{2}$$

The EC introduced by Bonacich [23, 24] is defined as the principal eigenvector of the adjacency matrix **A** defining the network. It simulates a mechanism in which each node affects all of its neighbors simultaneously. The defining equation of an eigenvector is $\lambda\mathbf{e} = \mathbf{Ae}$, where **A** is the adjacency matrix of the graph, λ is an eigenvalue and **e** is the eigenvector. Thus, EC of protein i is defined as the ith component of the eigenvector $\mathbf{e}_1$, $e_1(i)$, that corresponds to the largest eigenvalue of **A**, λ_1 (principal eigenvalue or index):

$$EC(i) = e_1(i) \tag{3}$$

Accordingly, a protein is considered central if it has a high eigenvector score, which means that it is adjacent to other proteins that themselves have high scores.

Another structural measure of centrality in a network was introduced by Stephenson and Zelen [25] and is known as IC. It is based on the information that can be transmitted between any two points in a connected network. If **A** is the adjacency matrix of a network, **D** a diagonal matrix of the degree of each node and **J** a matrix with all its elements equal to one, then IC is defined by inverting the matrix **B** defined as $\mathbf{B} = \mathbf{D} - \mathbf{A} + \mathbf{J}$ to obtain the matrix $\mathbf{C} = (c_{ij}) = \mathbf{B}^{-1}$ from which the information matrix is obtained as follows:

$$\mathbf{I}_{ij} = (c_{ii} + c_{jj} - c_{ij})^{-1} \tag{4}$$

The IC of the protein i is then defined by using the harmonic average:

$$IC(i) = \left[\frac{1}{N}\sum_{j}\frac{1}{\mathbf{I}_{ij}}\right]^{-1} \tag{5}$$

Stephenson and Zelen [25] proposed to define $\mathbf{I}_{ii}$ as infinite for computational purposes, which makes $1/\mathbf{I}_{ii} = 0$. Newman [30] has recognized that IC is another closeness measure, which in essence measures the harmonic mean lengths of paths ending at a node i, which is smaller if i has many short paths connecting it to other nodes.

Finally, we will consider a centrality measure recently introduced to account for the participation of a node in all subgraphs of the network. The SC is defined [26] as:

$$SC(i) = \sum_{l=0}^{\infty} \frac{\mu_l(i)}{l!} = \sum_{j=1}^{N} \left[\nu_j(i)\right]^2 e^{\lambda_j} \quad (6)$$

where $\mu_l(i)$ is the number of walks starting and ending at node *i*, *i.e.*, closed walks of length l starting at *i*, $(\nu_1, \nu_2, \ldots, \nu_n)$ is an orthonormal basis of R^N composed by eigenvectors of **A** associated to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, and $\nu_j(i)$ is the *i*th component of ν_j.

Accordingly, $SC(i)$ counts the total number of closed walks in which protein *i* takes part in the PIN and gives more weight to closed walks of short lengths. Closed walks are related to the network subgraph [26, 31]. Thus, SC accounts for the number of subgraphs in which a protein participates, giving more weights to smaller subgraphs, which have been previously identified as important structural motifs in biological networks [32–34].

2.2 Yeast PIN data set

The PIN of yeast, *S. cerevisiae*, used here, was compiled by Bu *et al.* [35]. The original data was obtained by von Mering *et al.* [36] by assessing a total of 80 000 interactions among 5400 proteins reported previously and assigning each interaction a confidence level. Bu *et al.* [35] focused on 11 855 interactions between 2617 proteins with high and medium confidence to reduce the interference of false positives, from which they reported a network consisting on 2361 nodes and 6646 links (http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm). This interaction map is considered here as a network in which proteins are represented as the nodes, and two nodes are linked by an edge if the corresponding two proteins can be expected with high or medium confidence of interacting. The main cluster of 2224 proteins sharing 6608 interactions, which is used here for the analysis, is illustrated in Fig. 1.

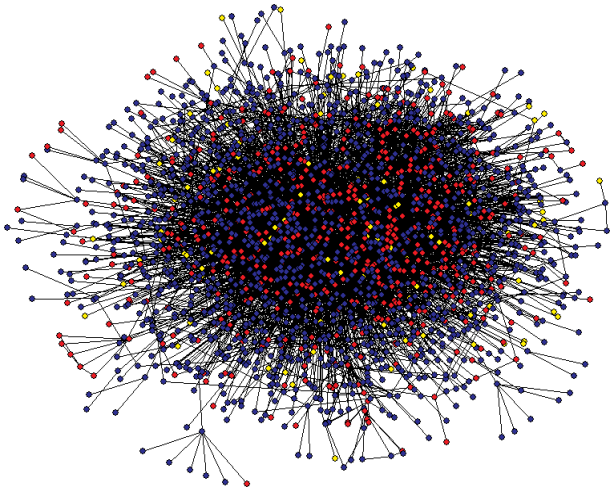

Figure 1. The complete PIN of yeast showing essential proteins in red and nonessential ones in blue, yellow circles correspond to proteins with unknown essentiality.

The indispensability of a protein defines the functional significance of a gene at its most basic level. Essential genes are those upon which the cell depends for its viability. Lethality can be determined without knowing the function of a gene by using, for example, random transposon mutagenesis [37] or gene-deletion [38]. Using the GENECENSUS database (http://bioinfo.mbb.yale.edu/genome/), we checked for all proteins in the main cluster of the yeast PIN for indispensability. Almost 80% of essential proteins in this main component of the yeast PIN form a connected cluster showing a high level of connectivity among essential proteins (see Fig. 2).

3 Results

The main objective of this work was to analyze the potential of centrality measures to select essential proteins in a PIN. DC was previously found to be correlated with protein lethality in yeast PINs in such a way that there is a larger number of essential proteins among those highly connected than among proteins with low degree [14]. Here our approach consisted in ranking proteins according to their values of centrality; we selected the top 1%, top 5%, etc. of proteins, and determined how many of these are essential in the yeast PIN. For purposes of comparison, we also selected proteins at random taking the average of the number of essential proteins after 20 random selections. In Fig. 3 we illustrate the number of essential proteins detected by SC, DC, CC, BC, EC and IC as well as by the random selection method up to the top 25% of proteins in the PIN, which represents an amount equivalent to the total number of essential proteins in the main cluster of the yeast PIN.

As can be seen in Fig. 3, all centrality measures perform significantly better than the random selection method in selecting essential proteins in the yeast PIN. This result is important from the practical point of view because it means that the use of centrality measures for selecting essential

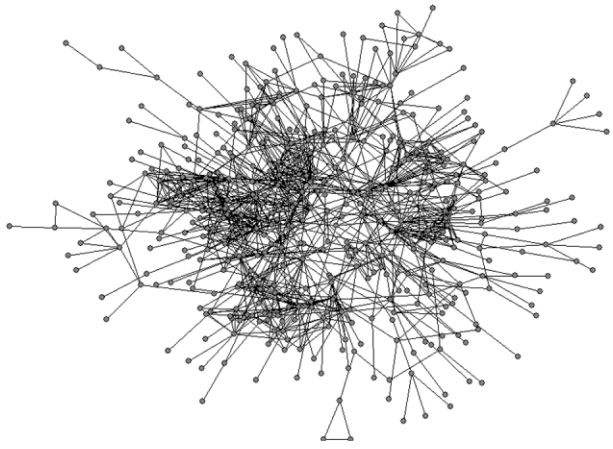

Figure 2. The main cluster of essential proteins of yeast containing 78.2% of all essential proteins in the PIN.

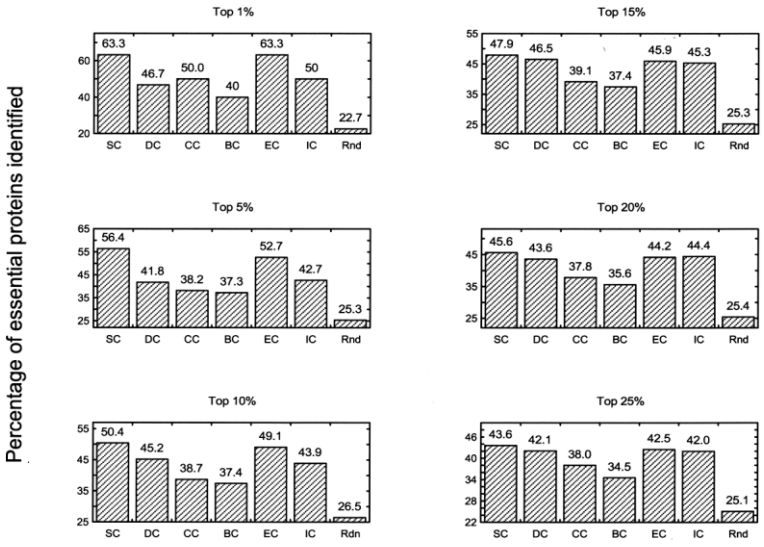

Figure 3. Percentage of essential proteins selected by ranking proteins according to their values of centrality and at random (after 20 realizations).

proteins in a PIN is significantly more efficient than random selection. If we select these proteins at random, we have a high chance of detecting a low ratio of essential proteins, and losing time, efforts and resources. However, the use of centrality measures to rank the proteins in this PIN and selecting, for instance, the top 1% of them will guarantee a higher probability of selecting essential proteins.

The percentage of essential proteins identified by SC show certain linear correlation with the percentages of essential proteins detected by the other centrality measures. These correlations are expected from the fact that all centrality measures are trying to identify the most central nodes in the network, even if they are using different criteria [39]. In particular, the best correlation was obtained for the EC, which is also a spectral centrality measure. In Fig. 4, we illustrate these correlations where it is also seen that all centrality measures outperform the random selection method.

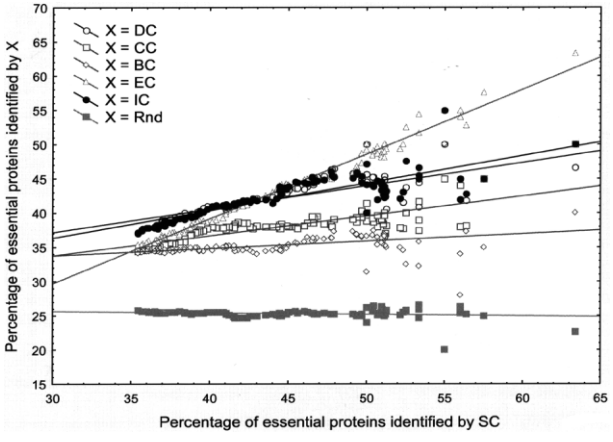

Figure 4. Linear correlations between the percentages of essential proteins selected by ranking proteins according to their values of SC versus those obtained by other centrality measures and at random (after 20 realizations).

We have tried to increase the percentage of essential proteins correctly identified by combining different centrality measures. For instance, we have selected the top 1% of proteins according to more than one centrality measure, and calculated the fraction of essential proteins existing among them. However, these percentages do not differ significantly from those obtained using SC. This is probably a consequence of the interrelation between the different centrality measures. Our hypothesis is that the essentiality of a protein in a proteome does not depends only on its topological position in the network but also on proper characteristics of the protein. Consequently, to increase the performance of a method for identifying essential proteins, we need to use a combination of topological and "molecular" information in the form of weighted networks.

An individual analysis of each centrality measure reveals interesting characteristics of this selection parameter. In most of the cases the SC measure identifies the larger number of essential proteins in the PIN among all centrality measures, followed very closely by the other spectral centrality measure, EC. The poorest performances are obtained by using BC and CC, while DC and IC report numbers of essential proteins intermediate between those of SC-EC and BC-CC. Any centrality measure is capable of identifying almost twice as many indispensable proteins as the random selection. If we consider the top 1% of proteins in the PIN to be comprised of 30 proteins, SC is able to identify 63.3% of these proteins as essential, 17% more than identified by DC. These differences are less marked as the percentage of proteins selected increases. For instance, when the top 25% of proteins are selected, SC identifies 8 essential proteins more than DC and only 6 more than EC. If we are interested in selecting a few proteins to test their lethality, so as to select some of them as targets for new drugs, the analysis of the top 1% of proteins ranked by centrality measures will be more appropriate than analysis of the top 25%. However, the dif-

ference with respect to BC and CC is still significant because SC identifies 51 essential proteins more than BC, and 31 more than CC. At this point SC overtakes the random selection method for more than 100 essential proteins.

The proteins selected by each of the centrality measures studied together with the links between them can be considered as sub-networks of the yeast PIN. For instance, if we take the top 1% proteins selected by any centrality measure we can form a network of 30 nodes together with the links joining them. These networks are not necessarily connected. Suppose that the top x nodes ranked by a centrality measure are topologically distant from each other with no link joining any pair of such nodes. In this case we will have a network formed by x disconnected nodes. However, if a link exists between each pair of these nodes the resulting sub-network will be connected. The study of the topological characteristics of these sub-networks can shed some light on the place that essential proteins occupy in the yeast PIN. In Table 1 we give the values of three topological characteristics for these sub-networks formed by the nodes ranked as the top 1%, 10% and 25% according to the different centrality measures as well as the random selection method. These topological parameters are the Watts and Strogatz [10] clustering coefficient, C; the average distance among reachable pairs of nodes, L, *i.e.*, the number of edges in the shortest path between each pair of nodes; and the average degree $<k>$.

Table 1. Topological characteristics of the sub-network formed by the top 25% of proteins selected by centrality measures or at random

Method		C	L[a)]	$<k>$
SC	1%	0.529	1.586	12.67
	10%	0.391	2.418	11.84
	25%	0.217	3.052	10.44
DC	1%	0.306	3.203	3.00
	10%	0.253	2.905	7.49
	25%	0.200	3.140	11.05
BC	1%	0.032	3.120	2.47
	10%	0.130	2.902	6.05
	25%	0.112	3.310	9.18
CC	1%	0.277	2.568	4.60
	10%	0.132	2.498	8.78
	25%	0.142	2.987	10.74
EC	1%	0.529	1.586	12.67
	10%	0.330	2.210	12.03
	25%	0.210	3.052	10.27
IC	1%	0.361	3.066	4.06
	10%	0.243	2.681	8.93
	25%	0.198	3.103	11.34
Random	1%	0.000	–[b)]	0.00
	10%	0.000	–[b)]	0.22
	25%	0.146	5.867	1.42

a) Average distance among reachable pairs of nodes.
b) Very low number of connected nodes.

The sub-networks formed by nodes selected by the two spectral methods show larger clustering coefficients and lower average distances among all the sub-networks studied. As expected, the sub-networks formed by random selection of nodes are highly disconnected, showing very low clustering. The clustering coefficient measures the degree of cliquishness of a node in the network and the sub-networks formed by selecting nodes according to their higher SC and EC show the largest cliquishness of all the sub-networks studied. Cliquishness refers to the graph theory term "cliquish" which is used to designate complete graphs, *i.e.*, those in which all nodes are connected to each other. Triangles, which are accounted by the clustering coefficient, are the complete graph with three nodes. Coincidentally, these two selection methods are the most efficient for the identification of essential proteins in the yeast PIN.

4 Discussion

The fact that centrality measures overtake the random selection in detecting essential proteins clearly indicate that these topological measures encode important structural information related to protein lethality, *i.e.*, they are describing some fundamental topological information of the PIN, which is likely to be essential for protein function. Our results also indicates that protein indispensability does not depend on how close a protein is to many other proteins, nor on the number of pairs of proteins that a particular protein needs as intermediary in their communications in the protein-protein interactions. More importantly, the protein essentiality appears to be related on how much a protein is implicated in clusters of proteins forming a large number of subgraphs in the network. This hypothesis is confirmed in part by the fact that the proteins selected by any of the spectral measures of centrality form clusters of highly interconnected nodes showing a high number of triangles as measured by the clustering coefficient. These results agree with those reported by Yu *et al.* [15], who determined that within the interaction network, essential proteins tend to be more cliquish.

Among these subgraphs, triangles and squares can play an important role in an understanding of the evolution of the PIN [32–34]. According to the coupled duplication-divergence model of evolution after gene duplication, both of the expressed proteins will have the same interactions [40]. In this model, it is proposed that both duplicate genes are subject to degenerative mutations, losing some functions but jointly retaining the full set of functions present in the ancestral gene. More recently, van Noort *et al.* [41] have reproduced the scale-free and small-world characteristics of the yeast co-expression network using a similar model, based on the simple neutralist's model, which consists of co-duplication of genes with their transcription factor binding sites (TFBSs), deletion and duplication of individual TFBSs, and gene loss [41]. Among the effects manifested by these models on the topology of the PIN is the tendency to generate bi-

connected triplets and quadruples of nodes; *i.e.*, triangles and squares. Triangles are formed among the duplicating genes and any neighbor of the parent gene, and squares are formed analogously between duplicating genes and any pair of neighbors of the parent gene. Triangles and squares can also be involved in a great number of other subgraphs, which are all well accounted for by SC. Consequently, SC shows the best performance in selecting essential proteins in the yeast PIN, probably because the indispensability of a given protein in the PIN is related to its implications in certain essential structural motifs formed by different structural subgraphs.

In conclusion, we have shown that the lethality associated with removal of a protein from the yeast proteome correlates with different centrality measures of the nodes in the PIN. These measures, which reflect different topological characteristics of networks, are significantly better than random selection in identifying essential proteins in a PIN. Our results confirm that the indispensability of a given protein in a PIN not only depends on the individual biochemical function and genetic redundancy of the protein, but also on the global organization of interactions and the topological role of such proteins in the network [14].

The centrality measures based on graph spectral properties of the network, in particular the SC, show the best performance in identifying essential proteins in the yeast PIN. The SC quantifies the participation of a node in the different subgraphs of the network, assigning more importance to the smaller ones, which have been identified as important motifs in biological networks [32–34]. Consequently, the use of SC in studying PINs can provide important structural information about the role of individual proteins in the global organization of protein interactions. At the same time ranking proteins by means of this centrality measure can be an effective method for selecting possible targets for rational drug discovery through the identification of essential proteins in a PIN.

5 References

[1] Barabási, A.-L., Oltvai, Z. N., *Nat. Rev. Genet.* 2004, *5*, 101–114.

[2] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabási, A.-L., *Nature* 2000, *407*, 651–654.

[3] Fell, D. A., Wagner, A., *Nat. Biotechnol.* 2000, *189*, 1121–1122.

[4] Koonin, E. V., Wolf, Y. I., Karev, G. P., *Nature* 2002, *420*, 218–223.

[5] Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K. *et al.*, *Curr. Opin. Struct. Biol.* 2004, *14*, 292–299.

[6] Vázquez, A., Flammini, A., Maritan, A., Vespignani, A., *ComplexUs* 2003, *1*, 38–44.

[7] Strogatz, S. H., *Nature* 2001, *410*, 268–276.

[8] Albert, R., Barabási, A.-L., *Rev. Mod. Phys.* 2002, *74*, 47–97.

[9] Newman, M. J. E., *SIAM Rev.* 2003, *45*, 167–256.

[10] Watts, D. J., Strogatz, S. H., *Nature* 1998, *393*, 440–442.

[11] Barabási, A.-L., Albert, R., *Science* 1999, *286*, 509–512.

[12] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., Barabais, A.-L., *Science* 2002, *297*, 1551–1555.

[13] Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S. *et al.*, *Nature* 2004, *430*, 88–93.

[14] Jeong, H., Mason, S. P., Barabási, A.-L., Oltvai, Z. N., *Nature* 2001, *411*, 41–42.

[15] Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., Gerstein, M., *Trends Genet.* 2004, *20*, 227–231.

[16] Vázquez, A., Flammini, A., Maritan, A., Vespignani, A., *Nat. Biotechnol.* 2003, *21*, 697–700.

[17] Pereira-Leal, J. B., Enright, A. J., Ouzounis, C. A., *Proteins* 2004, *54*, 49–57.

[18] Wagner, A., *Mol. Biol. Evol.* 2001, *18*, 1283–1292.

[19] Wutchy, S., *Genome Res.* 2004, *14*, 1310–1314.

[20] Wasserman, S., Faust, K., *Social Network Analysis*, Cambridge University Press, Cambridge 1994.

[21] Gómez, D., González-Arangüena, E., Manuel, C., Owen, G. *et al.*, *Math. Soc. Sci.* 2003, *46*, 27–54.

[22] Freeman, L. C., *Social Networks* 1979, *1*, 215–239.

[23] Bonacich, P., *J. Math. Sociol.* 1972, *2*, 113–120.

[24] Bonacich, P., *Am. J. Sociol.* 1987, *92*, 1170–1182.

[25] Stephenson, K., Zelen, M., *Social Networks* 1989, *11*, 1–37.

[26] Estrada, E., Rodríguez-Velázquez, J. A., *Phys. Rev. E.* 2005, *71*, 056103.

[27] Wuchty, S., Stadler, P. F., *J. Theor. Biol.* 2003, *223*, 45–53.

[28] Ma, H.-W., Zeng, A.-P., *Bioinformatics* 2003, *19*, 1423–1430.

[29] Freeman, L. C., *Sociometry* 1977, *40*, 35–41.

[30] Newman, M. J. E., *Social Networks* 2005, *27*, 39–54.

[31] Biggs, N., *Algebraic Graph Theory*, 2nd Edn., Cambridge Mathematical Library, Cambridge 1994.

[32] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R. *et al.*, *Science* 2004, *303*, 1538–1542.

[33] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N. *et al.*, *Science* 2002, *298*, 824–827.

[34] Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S. *et al.*, *Proc. Natl. Acad. Sci. USA* 2004, *101*, 5934–5939.

[35] Bu, D., Zhao, Y., Cai, L., Xue, H. *et al.*, *Nucleic Acids Res.* 2003, *31*, 2443–2450.

[36] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, *Nature* 2002, *417*, 399–403.

[37] Ross-Macdonald, P., Coelho, P. S. R., Roemer, T., Agarwal, S. *et al.*, *Nature* 1999, *402*, 413–418.

[38] Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H. *et al.*, *Science* 1999, *285*, 901–906.

[39] Wuchty, S., Stadler, P. F., *J. Theor. Biol.* 2003, *223*, 45–53.

[40] Force, A., Lynch, M., Pickett, F. B., Amores, A. *et al.*, *Genetics* 1999, *151*, 1531–1545.

[41] van Noort, V., Snell, B., Huynen, M. A., *EMBO Rep.* 2004, *5*, 1–5.

