The connectivity of large genetic networks: design, history, or mere chemistry?

Invited contribution to "Power laws, scale-free networks and genome biology"

By Karev, G., Wolf, Y., Koonin, E. (eds.)

Andreas Wagner

University of New Mexico and The Santa Fe Institute

University of New Mexico
Department of Biology

167A Castetter Hall
Albuquerque, NM 817131-1091
Phone: +505-277-2021
FAX: +505-277-0304
Email: wagnera@unm.edu

Abstract

I review evolutionary explanations of broad-tailed connectivity or degree distributions observed in metabolic networks and protein interaction networks. Self-assembled chemical reaction networks show degree distributions similar to those observed for metabolic networks, which argues against the postulated role of natural selection in maintaining this degree distribution. In addition, metabolic networks contain traces of their ancient history in the form of highly connected metabolites. Similarly to the degree distribution of metabolic networks, that of protein interaction networks can be explained without resorting to natural selection on the network level. I present data suggesting that highly connected proteins are not distinguishably older than other proteins, and explain this finding with a simple model of how a protein's degree changes in evolutionary time.

Introduction

Graph representations of biological networks have become popular with the recent accumulation of functional genomic data on such networks. Graphs are mathematical objects consisting of nodes and edges connecting these nodes. The *degree* or *connectivity* *d* of a node is the number of edges emanating from it, or, equivalently, the number of its neighbors in the graph. Multiple biological networks show a connectivity or degree distribution that is broad-tailed and often consistent with a power-law. That is, when choosing a node from such a network at random, the probability *P(d)* that it has *d* interaction partners is proportional to $P(d) \propto d^{-\gamma}$, γ being some constant that is characteristic of the network. Most prominently, this holds for metabolic networks, whose nodes can be substrates, reactions, or both, depending on the network representation one chooses, and protein interaction networks, where two nodes (proteins) are connected if they interact physically inside the cell. Broad-tailed degree distributions have also been demonstrated for other cellular networks. [1-3]

The degree distribution of a genetic network can be viewed as a feature of an organism like any other feature. It raises the same basic question: Why this and not some other degree distribution? There are three possible answers. First, a network's degree distribution could be a mere consequence of chemistry, the chemistry of DNA, RNA, and proteins, and the patterns of molecular interactions this chemistry allows. This possibility may seem far-fetched, given that molecular networks have many biological functions which may constrain their structure. However, this possibility is not without precedent. An illustrative example exists on a lower level of biological organization, protein structure. The thousands of currently known protein structures have a highly skewed distribution. There is a small number of 'frequent' tertiary structures, such as the TIM-barrel or the Rossman fold, found in nucleotide-binding proteins. While these folds are small in number, many proteins adopt them. Conversely, the majority of tertiary structures are 'unifolds' that may have originated only once in evolution, and are adopted by few proteins. [4-6] Does this skewed distribution of protein structures contain important information about design principles of proteins? For instance, do frequent structures have superior properties that lead to their frequent occurrence in proteins? The likely answer is no. Similarly skewed distributions of structures – a small number of excessively frequent structures and a vast majority of rare structures – occur in simple models of protein folding, models where polymers composed of parts with properties similar to amino acids fold into three-dimensional structures. [7, 8] The distribution of protein structures may be a mere consequence of polymer chemistry.

The second possibility is that the degree distribution of genetic networks might somehow reflect their history, much like the jumble of streets in a medieval city reflects the city's growth over centuries. An important class of mathematical models, originally devised to explain power-law degree distributions in growing networks like the internet, do indeed link a network's history to its degree distribution. In their original and simplest incarnation, such models involve only two simple rules that change the structure of a network. [9] First, the network grows through addition of nodes. Second, newly added nodes connect to previously existing nodes, such that already highly connected nodes are more likely to receive a new connection than nodes of lesser connectivity. Over many cycles of node addition and linking to existing nodes, a power law degree distribution emerges. A great variety of variations to this model have been proposed (reviewed in ref. 10). They differ greatly in detail but retain in some way or another the rule that new connections preferably involve highly connected nodes. Importantly, most such models make a key prediction: Highly connected nodes are old nodes, nodes having been added very early in a network's history. In this sense, they link a network' degree distribution to its history.

The third possibility is that molecular networks have their degree distribution, because this structure is somehow best suited to the network's biological function. From an 'organismal design' perspective, this is the most interesting possibility. It means that natural selection has shaped the global connectivity pattern of a network, and that network structure reveals something about the design principles of biological networks.

A recent hypothesis postulates that the observed broad-tailed degree distribution of biological networks is indeed a product of natural selection. [11-13] This 'selectionist' hypothesis is based on the following observation. In networks with a broad-tailed degree distribution, the mean distance between network nodes that can be reached from each other (via a path of edges) is very small and it increases only very little upon random removal of nodes. [11] (In contrast, this mean distance or mean path length increases drastically when highly connected nodes are removed.) A network's mean path length can be thought of as a measure of how 'compact' the network is. In graphs with other degree distributions, mean path length increases more substantially upon random node removal, and the network becomes more easily fragmented into disconnected components. These observations have led to the proposition that robustly compact networks confer some advantages on cells, and that a broad-tailed degree distribution reflects the action of natural selection on the degree distribution itself. The nature of this advantage is unknown, except in the case of metabolic networks, where one can venture an informed guess. [14, 15] A possible advantage of small mean path lengths in metabolic networks stems from the importance of minimizing transition times between metabolic states in response to environmental changes. [16-18] Networks with robustly small diameter may adjust more rapidly to environmental perturbations.

Metabolic networks and planetary atmospheres.

While the above speculation makes a weak case for a selectionist explanation of broad-tailed degree distributions in metabolic networks, another line of evidence makes a more solid case against it. One can ask whether power-law degree distributions might not be features of many or all large chemical reaction networks, whether or not part of an organism, whether or not they have a biological function which benefits from a robust network diameter. If so, then metabolic network degree distributions would join the club of other power-laws (such as Zipf's law of word frequency distributions in natural languages) whose existence does not owe credit to a benefit they provide. There is indeed evidence supporting this possibility.

Gleiss and collaborators [19] have compiled publicly available information on a class of large chemical reaction networks that exist not only outside the living, but on spatial scales many orders of magnitude larger than organisms. These are the chemical reaction networks of planetary atmospheres, networks whose structure is largely determined by the photochemistry of their component substrates. The available data stems not only from earth's atmosphere, but also from other solar planets including Venus and Jupiter, planets with chemically diverse atmospheres. These planets' atmospheres have been explored through remote spectroscopic sensing methods and by planetary probes. The chemical reaction networks in these atmospheres, despite being vastly different in chemistry, have a degree distribution consistent with a power law. [19] This suggests that power-law distributions may be very general features of chemical reaction networks. The reasons why we observe them in cellular reaction networks may have nothing to do with the robustness they may provide.

Although such comparisons to 'self-assembled' networks suggest an important influence of chemistry on metabolic network structure, another aspect of metabolic networks should not be overlooked. Metabolic networks have a history. They have not been assembled in their present state at once. They have grown, perhaps over a billion years, as organisms increased their metabolic and biosynthetic abilities. In understanding their structure, we have to take this history of biological networks into account.

We may never know enough about the history of life and metabolism to distinguish between different ways in which metabolism might have grown. However, we can address the key prediction of many network growth models I discussed above. *Are highly connected metabolites old metabolites?* The answer will contain a speculative element, because the oldest metabolites are those that arose in the earliest days of the living, close to life's origins. In addition, life forms as different as bacteria and humans have core metabolisms with a very similar structure. This suggests that the growth of metabolism has essentially been completed at the time the common ancestor of extant life emerged. Because this common ancestor does no longer exist, the detailed structure of its metabolism will remain in the dark forever. However, various hypotheses about life's origin make predictions on the chemical compounds expected to have been part of early organisms. There are several of these hypotheses, and they are complementary in the respect most important here: They emphasize the origins of different aspects of life's chemistry. Some emphasize the origins of the earliest genetic material, RNA. Others make postulates about the composition of the earliest proteins. Yet others ask about the earliest metabolites in energy metabolism. Each of them makes a statement about a different aspect of early life's chemistry.

Figure 1 shows the twelve most highly connected metabolites of the *E. coli* metabolic network graph. [14] Every single one of them has been part of early organisms according to at least one origin-of-life hypothesis. Colored in green are compounds such as coenzyme A thought to have been a part of early RNA-based organisms. [20] The RNA moieties such compounds contain are present in all organismal lineages. Some compounds in this group, such as tetrahydrofolate and coenzyme A, are thought to have played a role in precellular life that may have taken place on polykationic surfaces. These compounds are elongate molecules with one anionic terminus. They are therefore able to flexibly tether other molecules to the substrate, thus localizing them while simultaneously increasing their potential to react with other compounds. [21] Colored in red in Figure 1 are amino acids that were part of early proteins, based on likely scenarios for the early evolution of the genetic code. [22] Shown in blue are compounds likely to have been a part of early energy and biosynthetic metabolism. Glycolysis and the TCA cycle are perhaps the most ancient metabolic pathways, and various of their intermediates (α- ketoglutarate, succinate, pyruvate, 3-phosphoglycerate) occur in Figure 1 [20, 22-26]. The potential relation between evolutionary history and connectivity of metabolites corroborates a postulate put forth by Morowitz [23], namely that intermediary metabolism recapitulates the evolution of biochemistry.

In sum, the observation that power law degree distributions occur in self-assembled chemical reaction networks that were never under the influence of natural selection suggests that such distributions are a rather common feature of such networks. Natural selection on the level of this degree distribution is thus unnecessary to understand their origin. Metabolic networks have grown by addition of new metabolites, and their degree distribution is in tentative agreement with a general prediction of many network growth models: Highly connected metabolites tend to be phylogenetically old metabolites, metabolites that have been added very early in the evolution of metabolism.

Protein interaction networks

In contrast to chemical reaction networks, large and self-assembled protein interaction networks do not exist outside living cells. Thus, we can not hope to use arguments from self-assembled networks to argue for or against the role of natural selection in explaining a protein network's degree distribution. However, two different lines of evidence speak to this question for protein networks. The first class of evidence regards a corollary of the hypothesis that the degree distribution observed in genetic networks is a by-product of selection for 'robust compactness'. In networks with a broad-tailed degree distribution, mean path length increases drastically upon removal of highly connected nodes, as opposed to the removal of lowly connected nodes, which does not change dramatically. If it is network compactness that matters to the organism, then removal of highly connected nodes should have more severe effects on the fitness of the organism than removal of less highly connected nodes. This prediction of the selectionist hypothesis can be tested with a publicly available collection of yeast gene-knockout (synthetic-null) mutant strains. [27] Each strain of this collection lacks one gene, and the resulting change in growth rate has been measured under a variety of environmental conditions. [27-29] Jeong and collaborators

[13] first showed that a correlation between the effect of a gene-knockout mutation and the encoded protein's degree exists. Figure 2 illustrates this correlation with more recent data. [28]

The interpretation of data like that shown in Figure 2 faces multiple problems, aside from the fact that the association between protein degree and mutational effect is weak. The first problem is conceptual. While removal of highly connected proteins may have more severe effects on a cell, the reasons might have nothing to do with an altered network topology. For example, high connectedness may simple be an indicator that a protein acts in a variety of different cellular processes, hence the more severe defect when the protein is eliminated from a cell. Other problems in interpreting associations like that shown in Figure 2 are technical. First, the resolution at which the effect of a gene knock-out mutation on growth rate can be measured is very low. Much smaller fitness differences between wild-type and mutant cells than one can observe in the laboratory may lead to elimination of a mutant in the wild. Second, gene knock-out effects are usually measured only in one or a few laboratory environments, not in the myriad of conditions in which they could manifest themselves in the wild. Third, laboratory assays of gene knock-out effects usually measure only one or a few components of fitness – most prominently growth rate – and leave others, such as cell survival under starvation untouched. Because of these problems, it is not clear whether laboratory gene knock-out experiments measure quantities that reliably indicate the effects of such mutations on an organism's ability to survive and reproduce.

These technical problems – but not the previous, conceptual one – could be overcome with an evolutionary approach. Here, one assesses not gene knockout effects but the rate at which different proteins in a protein interaction network evolve. Specifically, one asks whether highly connected proteins have evolved more slowly than lowly connected proteins. If this is the case, then one can argue that their evolution is more severely constrained. Several pertinent studies are available. [30-33] Their results differ in details, partly because they are sensitive to which of several available protein interaction data sets one uses. [30] However, their main conclusion is the same. If there are differences in the evolutionary rates of proteins in a network, they are not due to the differential effects these proteins have on a network's compactness. Thus, evolutionary studies do not support the notion that natural selection for robust compactness is responsible for the broad-tailed degree distribution of protein interaction networks.

A completely different approach to testing the selectionist hypothesis is encapsulated in the following question. *Can we explain the structure of protein interaction networks from processes of molecular evolution whose rates we can estimate, without resorting to natural selection acting on the network as a whole*? The answer is yes. [34] Such an explanation may still involve natural selection, but on a *local* instead of a *global* scale. For example, whenever a mutation causes a new interaction between two proteins to occur, natural selection may determine whether this interaction becomes fixed in a population or eliminated from it, depending on whether the interaction is beneficial, neutral, or deleterious. However, this is selection acting on individual interactions rather than global properties of an entire network.

In a previous contribution, I have proposed an explanation of the protein interaction network's degree distribution from purely local processes such as gene duplications and mutations that generate new interactions and cause others to disappear.

[34] The rate at which some of these processes occur can be roughly estimated from available protein interaction data, and based on these estimates, one can establish a quantitative mathematical model that explains the network's structure. This explanation falls within a class of models for network evolution that involve preferential attachment, that is, highly connected proteins are more likely to evolve new interactions than other proteins. Empirical data supports the notion that preferential attachment occurs in protein interaction networks, as shown in Figure 3. Others have also proposed models of protein network evolution [35], models that differ in important details but that have one key commonality: They do not require natural selection on a global network feature, but they explain the network's structure from evolutionary events on the small, local scale of individual proteins.

Many models of network evolution based on preferential attachment predict that highly connected network nodes should be old nodes, nodes that were added very early in a network's history. [36] They should have arisen early in the evolution of the network. Because the protein interaction network shows preferential attachment (Fig. 3), the question arises whether such an association between protein age and connectivity exists. Specifically, one can ask whether highly connected proteins are phylogenetically old. Phylogenetically old proteins should have a wider taxonomic distribution than more recently evolved proteins. In two complementary analyses, I thus asked whether highly connected proteins have a wider phylogenetic distribution than less highly connected proteins.

Connectivity and protein age

For the first of these analyses, I used the fully sequenced genomes of six maximally diverse species. They represent fungi (*Schizosaccharomyces pombe*), protists (*Plasmodium falciparum*), plants (*Arabidopsis thaliana*), animals (*Drosophila melanogaster*), eubacteria (*Escherichia coli*), and archaea (*Methanococcus janaschii*). For each of the proteins in the protein interaction network of baker's yeast (*Saccharomyces cerevisiae*) I used gapped BLAST [37] to ask how many of these six species contain a recognizable homologue of the yeast proteins. The data in Fig. 4 show the results of this analysis for a BLAST protein alignment score threshold of $E<10^{-5}$ to identify homology**.** Specifically, the figure shows the average number of taxa that contain at least one homologue to a yeast protein (vertical axis) plotted against the degree of this protein in the protein interaction network. The analysis shown is based on two different data sets of yeast protein interactions. [38, 39] If highly connected proteins are phylogenetically old, then highly connected proteins should occur in significantly more of the six taxa than lowly connected proteins. The data of Figure 4, however, does not support this pattern. Figure 5 shows a complementary analysis, where I plotted average protein degree against the number of the six taxa in which a protein's homologue is found. If more widely distributed proteins are more highly connected, then they should have a higher degree. The data does not support this association either. Alignment score thresholds of $E<10^{-2}$ and $E<10^{-10}$ yield the same conclusion (data not shown).

In a second analysis, I cast my net wider than just the above six fully sequenced genome. I arbitrarily chose 15 highly connected proteins (degree > 4) and 15 proteins with low connectivity (degree one) from the yeast protein interaction network. [38] For each of these thirty proteins, I asked whether it has at least one homologue in any of six broad taxonomic groups: metazoa, plants, protists, fungi (exclusive *Saccharomyces* spp.), eubacteria, and archaea. Table 1 summarizes the results. Seven out of 15 highly connected proteins and six out of 15 proteins with degree one have homologues in all eukaryotes. The same proportion (12 out of 15) of highly connected proteins and proteins with degree one have homologues in fungi outside the genus *Saccharomyces*. The same holds also for proteins that have no homologues outside this genus (3 out of 15 proteins). Based on this data, it appears that highly connected yeast proteins are not phylogenetically older than proteins of low degree.

While this finding is at first sight puzzling, the following analysis suggests a mundane explanation. This explanation emerges from a stochastic model of how the number of a protein's interaction partners changes over time. Consider one protein in a protein interaction network and denote as D_t the number of proteins this protein interacts with. If time t is measured in suitable discrete units, such as million years, then the change of this variable over time can be represented by a first order Markov process. [40] Specifically, designate as p_i the probability that the protein gains an interaction, that is, that its degree increases by one (through a mutation that has become fixed in a population). Formally p_i=Prob$(D_t=i+1|D_{t-1}=i)$. Similarly, denote the probability that the protein loses an interaction by q_i (q_i=Prob$(D_t=i-1|D_{t-1}=i)$). Finally, let r_i denote the probability that D_t does not change between $t-1$ and t. This simple framework can capture a variety of observations. For instance, in an earlier contribution I suggested that the rate at which interactions get added and eliminated from the network must be approximately balanced, because of the high observed rate of interaction turnover. [34] This translates into $p_i \approx q_i$ for all i.

In addition, the observation that proteins with more interaction partners show a greater turnover of interactions (Fig. 3) can be captured as a dependency of p_i on i, e.g., $p_i=i\times c$, where c is some constant.

A quantity of interest in this stochastic process is the expected waiting time until a protein first returns to the state $D_t=i$, i.e., m_i=**E**($T_i|D_0$=i), where **E** indicates the expected value of the random variable T_i:=min$\{t>0$: $D_t=i\}$, which measures the time until the protein first visits state i. For i=0, this expected time m_i is closely related to the residence time of a protein in the network, that is, the time during which a protein has a degree greater than zero. Quantities like m_i are difficult to calculate because we do not know how p_i, q_i and r_i depend on i, especially for large i. However, it is noteworthy that if the above assumptions held for arbitrarily large i, then this stochastic process would belong in the class of null-recurrent Markov processes, [41] whose expected waiting time to return to any state (not only i=0) is infinite, and can thus not be calculated. We can, however, calculate related quantities that may explain why highly connected proteins are not necessarily phylogenetically old. Consider a protein with degree 1. What is the expected time until such a protein loses this interaction – and thus ceases to be part of the network – assuming that this protein never attains a degree higher than one? If we denote as τ the random variable measuring this time, then its distribution is given by Prob($\tau=k$)= $q_1r_1^{k-1}$, which is essentially a geometric distribution. Its mean and variance are given by **E**(τ)=$q_1/(1-r_1)^2$, and Var(τ)=$r_1q_1/(1-r_1)^3$. Order-of-magnitude estimates for upper bounds on the probabilities p_1 and q_1 suggest that they are of the order of 6x10^{-4} per protein and million year.[34] Using these values, **E**(τ) calculates as 416 million years, and its standard deviation as 588 million years. In other words, even a protein of low degree that does not acquire any further interactions through mutations takes more than an expected 400 million years to lose its only interaction, with an enormous standard deviation. For proteins that acquire more interactions in the course of evolution, this expected time would be much larger. Considering the standard deviation in and by itself, it is then hardly surprising that we can not distinguish proteins of different degrees by their phylogenetic distribution. The time for which even low degree proteins reside in the network can vary over an enormous range, a range greater than the time elapsed since the Cambrian radiation. A statistical test could not distinguish between the age of high and low-connectivity proteins if their residence time in a network can vary so widely.

Conclusions

In sum, I have reviewed evidence pertaining to the hypothesis that natural selection acts on the global structure of cellular networks and is responsible for their broad-tailed degree distribution. While associations between gene knock-out effects and protein degree weakly support this hypothesis for protein interaction networks, evolutionary studies and explanations of network structure based on purely local processes argue against it. I showed that the great dispersion of time for which proteins may reside in a network can obscure expected differences in the taxonomic distribution of highly and lowly connected proteins. Similar to metabolic reaction networks, where chemistry itself is an important factor shaping a network's structure, the minor role for natural selection in optimizing a network's degree distribution suggests an important role for protein chemistry in determining this distribution. Which of a protein's chemical features, such as domain composition or surface properties, renders some proteins highly connected? What aspect of protein chemistry is responsible for the observation that highly connected proteins show a greater evolutionary turnover of interactions? The answers to these and other questions are contained in accumulating structural data on thousands of proteins.

Acknowledgments

I would like to thank the Santa Fe Institute for its continued support, as well as the NIH for its support through grant GM063882.

Figure and Table Captions

Fig. 1.: Highly connected metabolites in *Escherichia coli* are evolutionarily old. The list shows the 12 most highly connected metabolites in the *E. coli* core intermediary metabolic network. The numbers in parentheses show the degree (number of neighbors) of a metabolite in the substrate network as defined by Wagner and Fell. [14] Green indicates proposed remnants of a surface metabolism or an RNA world. Red indicates proposed early amino acids. Blue indicates proposed early metabolites (in the tricarboxylic acid cycle or glycolysis). The network was generated after the elimination of the compounds NAD, ATP, and their derivatives. These are even more highly connected than the compounds shown here. They are also evolutionarily ancient. See text for further details.

Fig. 2.: A weak but significant correlation between protein degree and gene knockout effect. Information on protein degrees shown here was obtained by pooling data from three independent sources, two large-scale protein interaction studies [38, 42], and a public data base of protein interactions [39] from which all interactions generated with the yeast two-hybrid assay had been eliminated. The horizontal axis shows the difference in the growth rate of a gene knock-out strain between the growth medium (among five different media) in which the strain grew at the highest rate, and the medium in which it grew at the lowest rate, as reported by Steinmetz and collaborators. [28] Growth rates are measured relative to a large pool of yeast gene deletion strains. [28]

For most genes, the growth rate difference is an indicator of the largest gene knockout effect among the tested growth media. An analogous analysis using the growth rate change of a gene knockout mutation in only rich medium (YPD) yields the same results (not shown).

Fig. 3.: Preferential attachment in protein interaction networks. The horizontal axis shows protein degree d. The vertical axis shows the likelihood P_d that a protein of degree d evolves new interactions. This likelihood can be estimated from the number of newly evolved interactions between products of paralogous genes, as detailed in ref. 34. For all member genes of a paralogous gene pair with a newly evolved interaction since their duplication, I determined the number I_d of those genes whose encoded proteins had d interactions to proteins different from its paralogue. To account for the fact that proteins of different degree occur at different frequencies in the network, I then divided this number by the relative frequency f_d of proteins of degree d in the network, and normalized the resulting quantity to obtain $P_{d,}$ i.e., $P_d=(I_d/f_d))/\Sigma_d(I_d/f_d)$. There is a strong, approximately linear association between protein degree and the likelihood to evolve new interactions. From Figure 5 in ref. 34.

Fig. 4.: The vertical axis shows the average number of genomes (± one s.d.) among six fully sequenced genomes that contain at least one protein homologous to proteins whose degree is indicated on the horizontal axis. The analysis is based on two different data sets on yeast protein interactions, one ('two hybrid') from a high-throughput experiment using

the yeast-two hybrid assay to identify such interactions [38], the other ('non-two hybrid') from a publicly available database on protein interactions from which I eliminated all data generated with the two-hybrid assay. [39] Protein comparisons are based on the following six maximally diverse fully sequenced and publicly available genomes: *Schizosaccharomyces pombe* (www.sanger.ac.uk), *Plasmodium falciparum* (www.plasmodb.org), *Arabidopsis thaliana* (www.tigr.org), *Drosophila melanogaster* (www.fruitfly.org), *Escherichia coli* K12-MG1655 (www.tigr.org), *Methanococcus janaschii* DSM2661 (www.tigr.org). I used gapped BLAST [37] with a threshold protein alignment score of $E<10^{-5}$ to identify homology. Results (not shown) are qualitatively identical for threshold scores of $E<10^{-2}$ and $E<10^{-10}$.

Fig. 5.: The vertical axis shows the average degree (± one s.d.) of proteins in the yeast protein interaction network as a function of the number of genomes – among six fully sequenced genomes – in which these proteins contain homologues, as shown on the horizontal axis. The analysis is based on two different data sets on yeast protein interactions, one ('two hybrid') using the yeast-two hybrid assay to identify such interactions [38], the other ('non-two hybrid') a publicly available database on protein interactions from which I eliminated all data generated with the two-hybrid assay. [39] Protein comparisons are based on the following six maximally diverse fully sequenced and publicly available genomes: *Schizosaccharomyces pombe* (www.sanger.ac.uk), *Plasmodium falciparum* (www.plasmodb.org), *Arabidopsis thaliana* (www.tigr.org), *Drosophila melanogaster* (www.fruitfly.org), *Escherichia coli* K12-MG1655 (www.tigr.org), *Methanococcus janaschii* DSM2661 (www.tigr.org). For the data shown, I used gapped BLAST [37] with a threshold protein alignment score of $E<10^{-5}$ to identify homology. Results (not shown) are qualitatively identical for threshold scores of $E<10^{-2}$ and $E<10^{-10}$.

Table 1: Taxonomic distribution of proteins with different connectivity in the yeast protein interaction network. The upper and lower parts of the table show the phylogenetic distribution of 15 arbitrarily chosen high and low degree proteins from publicly available yeast protein interaction data [38]. Gapped BLAST [37] was used to search for homologs to these yeast proteins in the GenBank database (www.ncbi.nlm.nih.gov). Columns in the table correspond to the following broad taxonomic groups. Metazoa (M), Protists (Pr), Plants (P), Fungi (F, exclusive of the genus Saccharomyces), Eubacteria (E) and Archaea (Ar). A '+' indicates that the respective protein has at least one putative homologue within the respective taxonomic group with a BLAST amino acid alignment score of $E<10^{-10}$. '++' and '+++' indicate at least one homologue with $E<10^{-20}$ and $E<10^{-30}$, respectively.

Twelve key metabolites in *E. coli* ranked by degree ("connectivity")

glutamate (51)
pyruvate (29)
coenzyme A (29)
α-ketoglutarate (27)
glutamine (22)
aspartate (20)
acetyl-CoA (17)
phosphoribosyl pyrophosphate (16)
tetrahydrofolate (15)
succinate (14)
3-phosphoglycerate (13)
serine (13)

Fig. 1

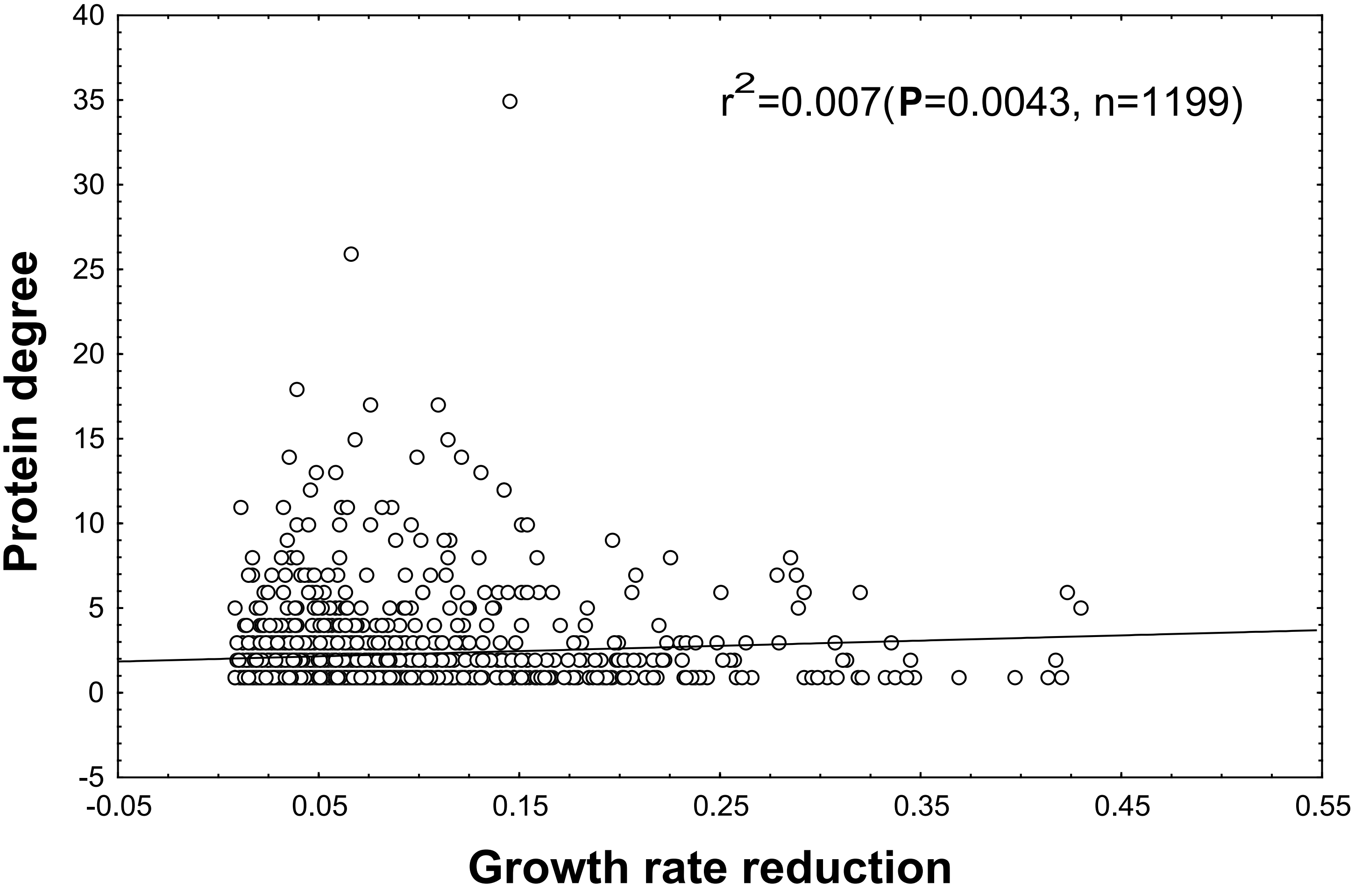
40
35
30
25
20
15
10
5
0
-5
-0.05
0.05
0.15
0.25
0.35
0.45
0.55
Protein degree
Growth rate reduction
r²=0.007(P=0.0043, n=1199)

Fig. 2

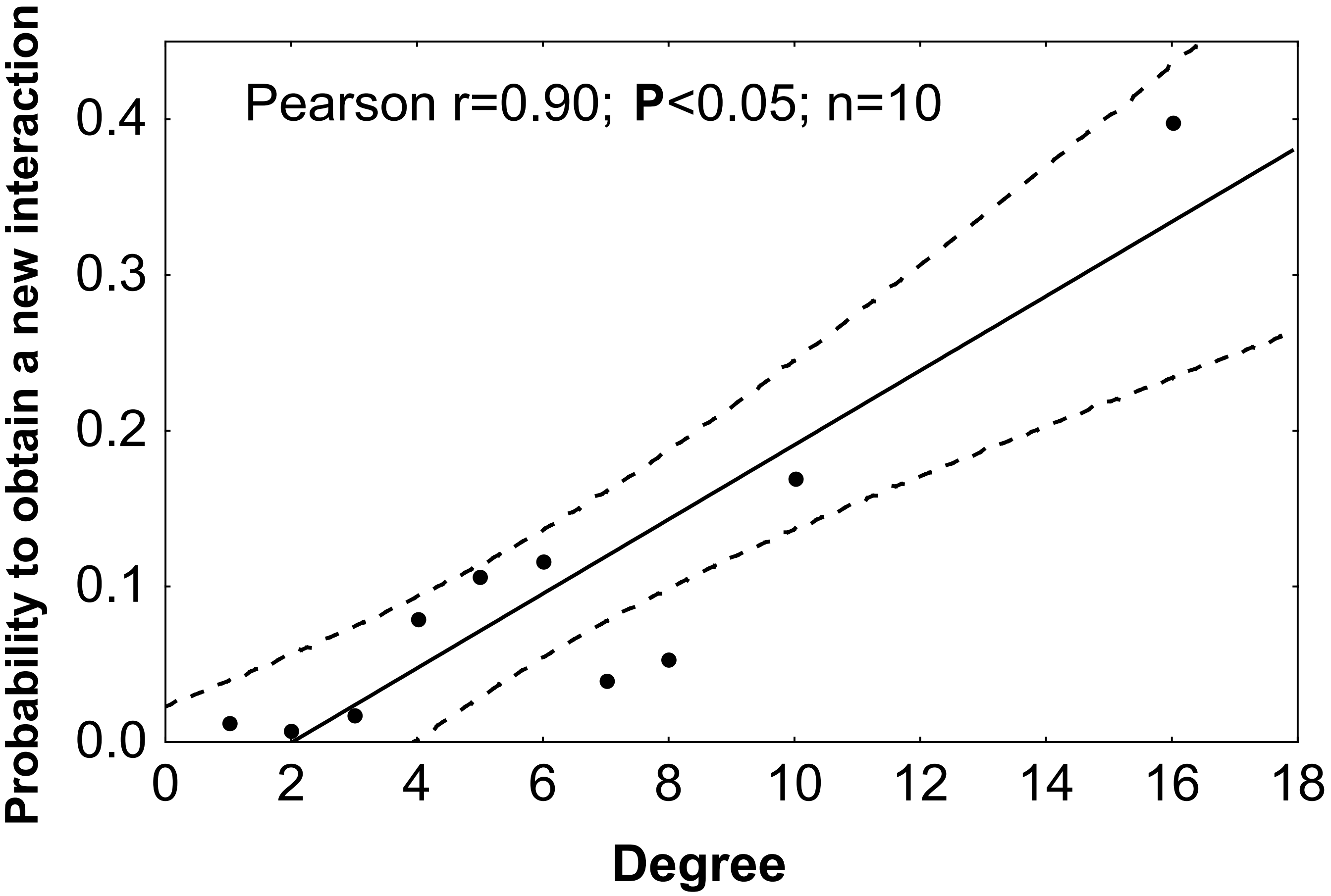
Pearson r=0.90; P<0.05; n=10
Probability to obtain a new interaction
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
12
14
16
18
Degree

Fig. 3

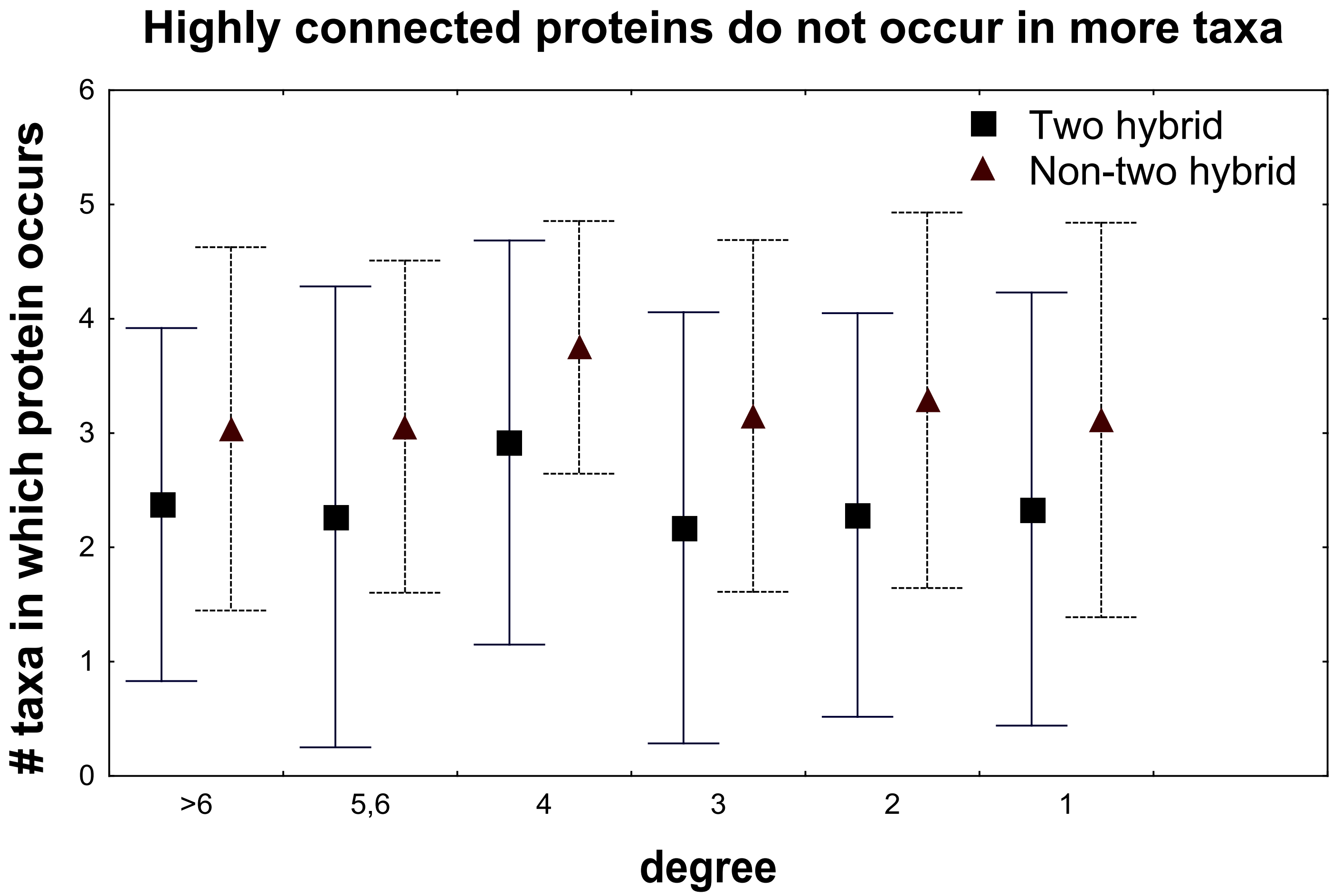
Highly connected proteins do not occur in more taxa
# taxa in which protein occurs
0
1
2
3
4
5
6
degree
>6
5,6
4
3
2
1
■ Two hybrid
▲ Non-two hybrid

Fig. 4

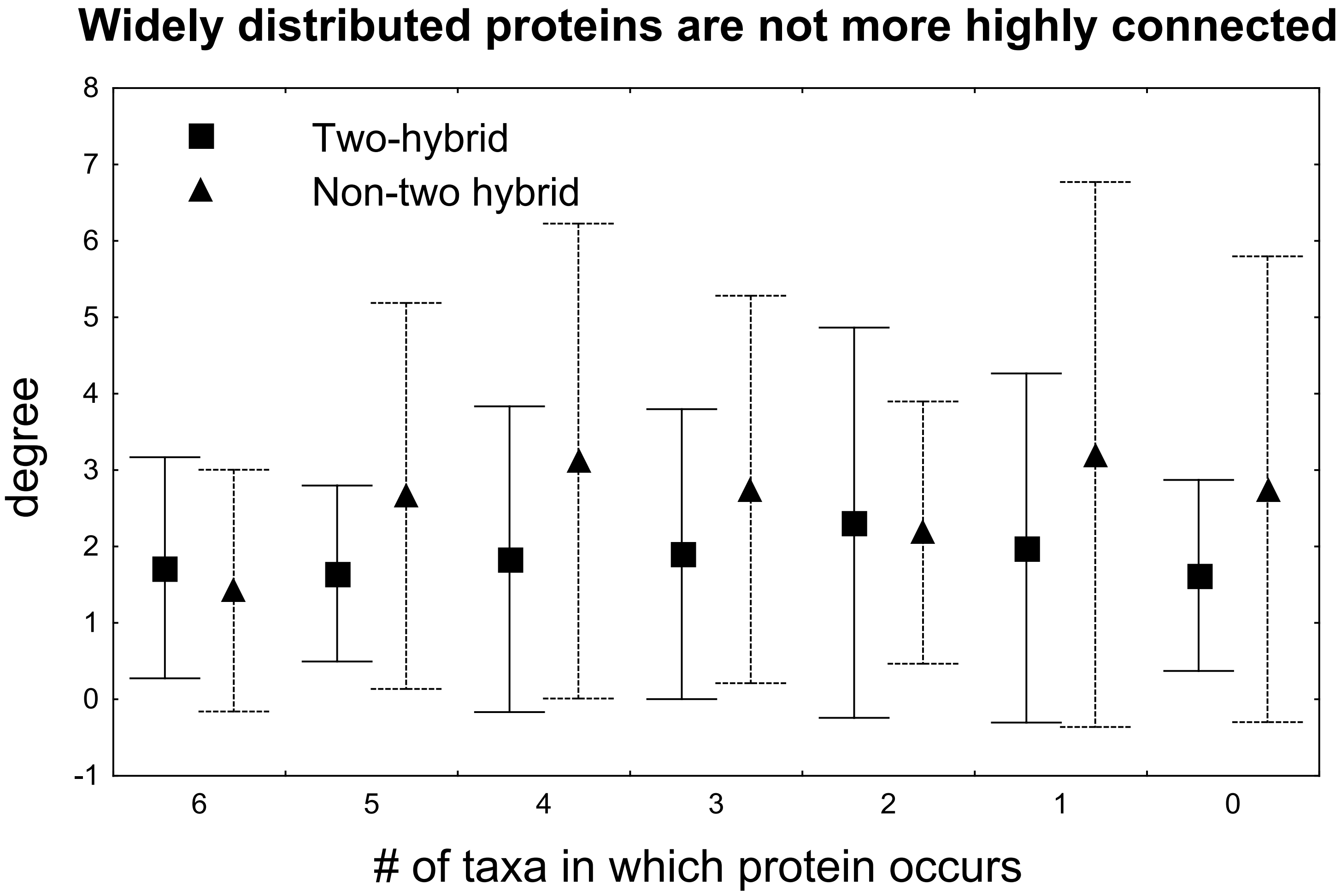
Widely distributed proteins are not more highly connected
■ Two-hybrid
▲ Non-two hybrid
degree
8
7
6
5
4
3
2
1
0
-1
6
5
4
3
2
1
0
# of taxa in which protein occurs

Fig. 5

Table 1

High Degree Proteins

Name	Deg	M	Pr	P	F	B	Ar
GLC1	12	+++	+++	+++	+++	-	-
CDC7	11	+++	+++	+++	+++	-	-
PHO85	10	+++	+++	+++	+++	++	-
LSM4	9	+++	-	+	+	-	-
SAP4	8	-	-	-	+++	-	-
CSM1	8	-	-	-	-	-	-
YCK2	7	+++	+++	+++	+++	+	-
YIL105C	7	-	-	-	++	-	-
MET30	7	+++	++	++	+++	++	-
YDL012C	7	-	-	-	-	-	-
CLB2	6	+++	++	+++	+++	-	-
CVT19 6	-	-	-	-	-	-	
ERF2	6	+++	+	+++	+++	-	-
CUP2	6	-	-	-	++	-	-
RPC19	5	++	-	-	++	-	-

Low Degree Proteins

Name	Degree	M	Pr	P	F	E	Ar
VPS4	1	++	+++	+++	+++	+++	+++
RHO1	1	++	+++	+++	+++	-	-
KRE6	1	-	-	-	+++	-	-
SMK1	1	++	+++	+++	+++	+	-
RLF2	1	-	-	-	-	-	-
YPR011C	1	+++	++	+++	++	-	-
YPR008W	1	-	-	-	++	-	-
APM1	1	+++	-	+++	+++	-	-
VIK1	1	-	-	-	-	-	-
HRR25	1	+++	+++	+++	+++	+	-
MKK2	1	+++	+++	+++	+++	-	-
YPL110C	1	+++	-	++	+++	-	-
MET31	1	-	-	-	-	-	-
YPL019C	1	-	++	-	+++	-	-
GDH1	1	-	+++	+++	+++	+++	-

Literature Cited

1. Rzhetsky A, Gomez SM. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics 2001;17(10):988-996.
2. Wuchty S. Scale-free behavior in protein domain networks. Molecular Biology and Evolution 2001;18(9):1694-1702.
3. Wuchty S. Interaction and domain networks of yeast. Proteomics 2002;2(12):1715-1723.
4. Koonin E, Wolf Y, Karev G. The structure of the protein universe and genome evolution. Nature 2002;420(6912):218-223.
5. Branden C, Tooze J. Introduction to protein structure New York: Garland; 1999.
6. Nagano N, Orengo C, Thornton J. One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. Journal of Molecular Biology 2002;321(5):741-765.
7. Li W-H. Molecular Evolution Massachusetts: Sinauer; 1997.
8. Bornberg-Bauer E. How are model protein structures distributed in sequence space? Biophysical Journal 1997;73(5):2393-2403.
9. Barabasi A-L, Albert, R, Jeong, H. Mean-field theory for scale-free random networks. Physica A 1999; 272(1-2): 173-187.
10. Albert, R., Barabasi, A-L. Statistical mechanics of complex networks. Reviews of Modern Physics 2002; 47(1): 47-94
11. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature 2000;406(6794):378-382.
12. Jeong H, Tombor, B. Albert, R. Oltvai, Z.N., Barabasi, A.L. The large-scale organization of metabolic networks. Nature 2000;407:651-654.
13. Jeong H, Mason SP, Barabasi A-L, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001;411:41-42.
14. Wagner A, Fell D. The small world inside large metabolic networks. Proc. Roy. Soc. London Ser. B 2001;280:1803-1810.
15. Fell D, Wagner A. The small world of metabolism. Nature Biotechnology 2000;18:1121-1122.
16. Cascante M, Melendez--Hevia E, Kholodenko BN, Sicilia J, Kacser H. Control analysis of transit--time for free and enzyme--bound metabolites - physiological and evolutionary significance of metabolic response--times. Biochemical Journal 1995;308:895-899.
17. Easterby JS. The effect of feedback on pathway transient response. Biochemical Journal 1986;233:871-875.
18. Schuster S, Heinrich R. Time Hierarchy in Enzymatic-Reaction Chains Resulting From Optimality Principles. Journal of Theoretical Biology 1987;129(2):189-209.
19. Gleiss PM, Stadler PF, Wagner A, Fell DA. Small cycles in small worlds. Advances in complex systems 2001;4:207-226.
20. Benner SA, Ellington AD, Tauer A. Modern metabolism as a palimpsest of the RNA world. Proceedings of the National Academy of Sciences of the U.S.A. 1989;86:7054-7058.

21. Wachtershauser G. Before enzymes and templates: theory of surface metabolism. Microbiological reviews 1988;52:452-484.
22. Kuhn H, Waser J. On the origin of the genetic code. FEBS letters 1994;352:259-264.
23. Morowitz HJ. Beginnings of cellular life New Haven: Yale University Press; 1992.
24. Taylor BL, Coates D. The code within the codons. Biosystems 1989;22:177-187.
25. Waddell TG, Bruce GK. A new theory on the origin and evolution of the citric acid cycle. Microbiologia sem 1995;11:243-250.
26. Lahav N. Biogenesis New York: Oxford University Press; 1999.
27. Giaever G, Chu AM, Ni L, et al. Functional profiling of the Saccharomyces cerevisiae genome. Nature 2002;418(6896):387-391.
28. Steinmetz L, Scharfe C, Deutschbauer A, et al. Systematic screen for human disease genes in yeast. Nature Genetics 2002;31(4):400-404.
29. Winzeler EA, Shoemaker DD, Astromoff A, et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 1999;285(#5429):901-906.
30. Hahn M, Conant GC, Wagner A. Molecular evolution in large genetic networks: connectivity does not equal importance. (in review) 2003.
31. Fraser HB, Wall DP, Hirsh AE. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evolutionary Biology 2003;3:11.
32. Jordan IK, Wolf YI, Koonin EV. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evolutionary Biology 2003;3:1.
33. Jordan IK, Wolf YI, Koonin EV. Correction: no simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors evolve slowly. BMC Evolutionary Biology 2003;3:5.
34. Wagner A. How large protein interaction networks evolve. Proceedings of the Royal Society of London Series B 2003;270:457-466.
35. Sole RV, Pastor-Satorras R, Smith ED, Kepler T. A model of large-scale proteome evolution. Advances in Complex Systems 2002;5: 43-54
36. Albert R, Barabasi AL. Statistical mechanics of complex networks. Reviews of Modern Physics 2002;74(1):47-97.
37. Altschul SF, Madden TL, Schaffer AA, et al. Gapped Blast and Psi-Blast : a new generation of protein database search programs. Nucleic Acids Research 1997;25(17):3389-3402.
38. Uetz P, Giot L, Cagney G, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000;403(6770):623-627.
39. Mewes HW, Heumann K, Kaps A, et al. MIPS: a database for genomes and protein sequences. Nucleic Acids Research 1999;27:44-48.
40. Karlin S. A first course in stochastic processes. New York: Academic Press, 1975.
41. Kulkarni VG. Modeling and analysis of stochastic systems. New York: Chapman & Hall, 1995.
42. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the

National Academy of Sciences of the United States of America 2001;98(8):4569-4574.