

DETECTION OF TOPOLOGICAL PATTERNS IN PROTEIN NETWORKS

Sergei Maslov and Kim Sneppen

December 30, 2003

Physics Department

Brookhaven National Laboratory
Operated by
Brookhaven Science Associates
Upton, NY 11973

Under Contract with the United States Department of Energy
Contract Number DE-AC02-98CH10886

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state to reflect those of the United States Government or any agency thereof.

Detection of topological patterns in protein networks

Sergei Maslov¹, Kim Sneppen²

Introduction

Complex networks appear in biology on many different levels:

- All biochemical reactions taking place in a single cell constitute its metabolic network, where nodes are individual metabolites, and edges are metabolic reactions converting them to each other.
- Virtually every one of these reactions is catalyzed by an enzyme and the specificity of this catalytic function is ensured by the key and lock principle of its physical interaction with the substrate. Often the functional enzyme is formed by several mutually interacting proteins. Thus the structure of the metabolic network is shaped by the network of physical interactions of cell's proteins with their substrates and each other.
- The abundance and the level of activity of each of the proteins in the physical interaction network in turn is controlled by the regulatory network of the cell. Such regulatory network includes all of the multiple mechanisms in which proteins in the cell control each other including transcriptional and translational regulation, regulation of mRNA editing and its transport out of the nucleus, specific targeting of individual proteins for degradation, modification of their activity e.g. by phosphorylation/dephosphorylation or allosteric regulation, etc. To get some idea about the complexity and interconnectedness of protein-protein regulations in baker's yeast *Saccharomyces Cerevisiae* in Fig. 1 we show a part of the regulatory network corresponding to positive or negative regulations that regulatory proteins exert on *each other*.

¹Department of Physics, Brookhaven National Laboratory, Upton, New York 11973, USA; E-mail: maslov@bnl.gov

²Nordita, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark; E-mail: sneppen@nbi.dk

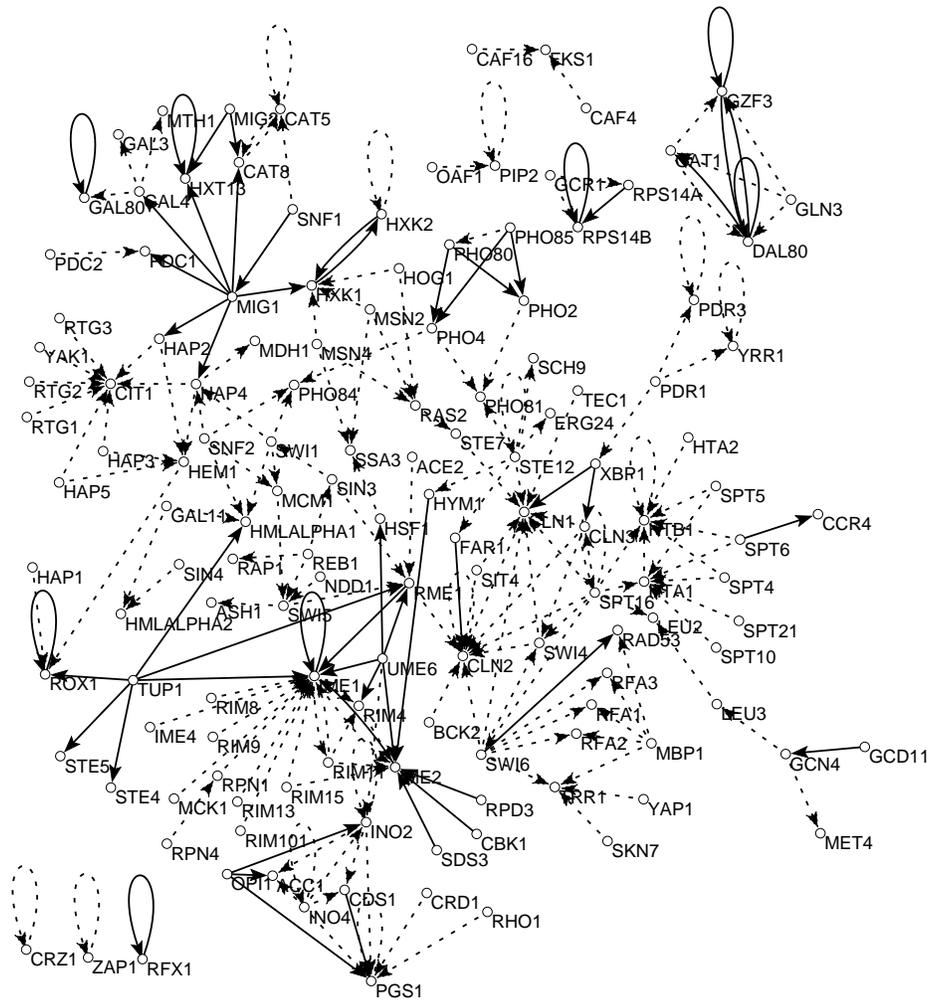


Figure 1: Regulations between regulators in yeast. Solid (dashed) arrows represent positive (negative) regulations of one yeast regulatory protein by another. The full list of known protein-protein regulations in yeast downloaded from the Yeast Proteome Database (YPD) [1] has 262 regulatory proteins and 1772 positive or negative regulations. This figure shows all 219 regulations that regulatory proteins exert on each other. The Pajek software [2] was used in drawing the network.

- On yet higher level individual cells of a multicellular organism exchange signals with each other. This gives rise to several new networks such as e.g. nervous, hormonal, and immune systems of animals. The inter-cellular signaling network stages the development of a multicellular organism from the fertilized egg.
- Finally, on the grandest scale, the interactions between individual species in ecosystems determine their food webs.

An interesting property of many biological networks that was recently brought to attention of the scientific community [3, 4, 5] is an extremely broad distribution of node connectivities defined as the number of immediate neighbors of a given node in the network. While the majority of nodes have just a few edges connecting them to other nodes in the network, there exist some nodes, that we will refer to as “hubs”, with an unusually large number of neighbors. The connectivity of the most connected hub in such a network is typically several orders of magnitude larger than the average connectivity in the network. Often the distribution of connectivities of individual nodes can be approximated by a scale-free power law form [3] in which case the network is referred to as scale-free. Among biological networks distributions of node connectivities in metabolic [4], protein interaction [5], and brain functional [6] networks can be reasonably approximated by a power law extending for several orders of magnitude.

The set of connectivities of individual nodes is an example of a low-level (single-node) topological property of a network. While it answers the question about how many neighbors a given node has, it gives no information about the identity of those neighbors. It is clear that most functional properties of networks are defined at a higher topological level in the exact pattern of connections of nodes to each other. However, such multi-node connectivity patterns are rather difficult to quantify and compare between networks.

In this work we concentrate on multi-node topological properties of protein networks. These networks (as any other biological networks) lack the top-down design. Instead, selective forces of biological evolution shape them from raw material provided by random events such as mutations within individual genes, and gene duplications. As a result their connections are characterized by a large degree of randomness. One may wonder which connectivity patterns are indeed random, while which arose due to the network growth, evolution, and/or its fundamental design principles and limitations?

In the next chapter we describe a universal recipe for how such information can be extracted. To this end one first constructs a proper randomized version (null model) of a given network. As was pointed out in the general context of complex scale-free networks [7], a broad distribution of connectivities indicates that the connectivity itself is an important individual characteristic of a node and as such it should be preserved in the randomized null-model network. In addition to connectivities one may choose to preserve some other low-level topological properties of the network in question. Any measurable topological quantity, such as e.g. the total number of edges connecting pairs of nodes with given connectivities, the number of loops of a certain type, the number and sizes of components, the diameter of the network, can then be measured in the real complex network and separately in its randomized version. One then concentrates only on those topological properties of the real network that significantly deviate from its null model counterpart.

The plan of this review is as follows: In the next chapter we introduce our basic algorithm for generation of an ensemble of randomized networks [8, 9]. We also propose a modification of this algorithm conserving other low-level topological properties of the network in addition to connectivities of its nodes [9]. In Chapter 3 we use these random ensembles to measure correlation profiles of several protein networks, namely those of physical interactions and transcriptional regulation between proteins in yeast *Saccharomyces Cerevisiae* [8]. Finally, the potential meaning of the observed large-scale properties of protein networks is discussed in Chapter 4. The set of MATLAB programs used to generate an ensemble of randomized networks, as well as to construct, and visualize correlation profiles of any given complex network can be downloaded from [10].

Local rewiring algorithm and topological profiles of a network

One may generate a random version of a given network using various algorithms. They differ from each other by which low-level topological features of the original network are preserved in its randomized counterpart. Below we list three such randomization processes in order of the increasing number of constraints:

1. Randomly rewire all edges in the network. This algorithm only conserves the *average* connectivity of all nodes in the network.
2. Randomly rewire edges in the network while preserving the number of edges emanating from each individual node (node's connectivity). This algorithm conserves all single-node topological properties of a network, while completely randomizes multi-node connection patterns. In a directed network one may rewire edges in such a way that both the number of outgoing and incoming edges are separately conserved for each node.
3. In addition to connectivities one may choose to conserve other low-level topological properties of the network such as e.g. the number of loops of a given type, the number and sizes of its components, its modular structure, etc.

The first rewiring scheme always generates an Erdős-Rényi (ER) random network [11] characterized by a narrow Poisson distribution $p(k) = \langle k \rangle^k \exp(-\langle k \rangle) / k!$ of node connectivities k , irrespective of the original form of this distribution. As topological properties of a network are very sensitive to the exact functional form of this distribution (in particular to its second moment) [12, 7], they typically would be modified as a result of the simple minded randomization algorithm #1. The change would be especially dramatic for networks with a broad (e.g. scale-free) distribution of connectivities. Therefore, for this class of networks a much more informative comparison would be to a randomized network generated by algorithms 2-3, where connectivities of individual nodes are strictly conserved.

The local rewiring algorithm giving rise to such random network was proposed in [13, 8]. It consists of multiple repetitions of the following simple switch move (elementary rewiring step) illustrated in Fig. 2:

Randomly select a pair of edges $A \rightarrow B$ and $C \rightarrow D$ and rewire them in such a way that A becomes connected to D , while C to B , provided that none of these new edges already exist in the network, in which case the rewiring step is aborted and a new pair of edges is selected.

The last restriction prevents the appearance of multiple edges connecting the same pair of nodes. A repeated application of the above rewiring step

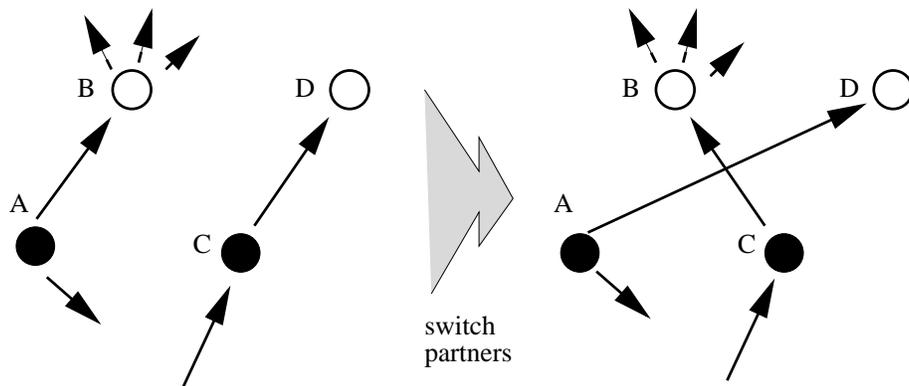


Figure 2: One step of the random local rewiring algorithm. A pair of edges $A \rightarrow B$ and $C \rightarrow D$ is randomly selected. The two edges are then rewired in such a way that A becomes connected to D , while C to B , provided that none of these new edges already exist in the network, in which case the rewiring step is aborted and a new pair of edges is selected. An independent random network is obtained when the network is rewired by the above local switch a large number of times, say several times in excess of the total number of edges in the system. Note that for directed networks this rewiring algorithm separately conserves both the in- and out- connectivities of each individual node.

leads to a randomized version of the original network. The set of MATLAB programs generating such a randomized version of any complex network can be downloaded from [10].

Sometimes it is desirable that the null-model random network in addition to nodes' connectivities conserves some other topological quantity of the real network. In this case one could still use the random rewiring algorithm, described above, but supplement it with the Metropolis acceptance/rejection criterion [14] of a simple switch move [8, 9].

For the sake of concreteness let's assume that one wants to generate a random network with the same set of nodes' connectivities and the same number N_Δ of triangles as the real undirected network [9]. Indeed, the number of triangles in a network is related to its "clustering coefficient" routinely used as a measure of modularity [15]. Hence, by conserving N_Δ one generates a null-model with the same average level of modularity as the original complex network.

The Metropolis version of the random rewiring algorithm uses an artificial energy function H that favors the number of triangles in a random network $N_\Delta^{(r)}$ to be as close as possible to its value N_Δ in the real network:

$$H = \frac{(N_\Delta^{(r)} - N_\Delta)^2}{N_\Delta} . \quad (1)$$

The Metropolis rules in this case allow for any local rewiring step that lowers the energy H or leaves it unchanged. However, those steps that lead to a ΔH increase in the "energy" H are accepted only with a probability $\exp(-\Delta H/T)$. Here the exact rules of the algorithm depend on (typically very small) "temperature" T introduced to prevent the sequence of rewiring steps from getting stuck in local (often suboptimal or non-representative) energy minima. In order to get a random network with $N_\Delta^{(r)}$ sufficiently close to N_Δ the temperature should be selected to be as small as possible without sacrificing the ergodicity of the problem. In the end one could always "prune" the resulting ensemble of random networks by leaving only networks with $N_\Delta^{(r)} = N_\Delta$.

The above Metropolis algorithm can be easily extended to take care of several independent topological motifs by using the composite energy function $H = \sum_m H_m$, where the index m runs over the desired set of motifs.

Once the desired null model random network is generated one could ask the question: which topological quantities in the real complex network significantly deviate from their values in its typical random counterpart. Such deviations can be quantified by the following set of network's *topological profiles*.

In the first profile one computes the ratio

$$R(j) = \frac{N(j)}{N_r(j)} \quad (2)$$

where $N(j)$ is the number of times the pattern j is seen in the real network, and $\overline{N_r(j)}$ is the average number of its occurrences in an ensemble of randomized networks, generated e.g. by one of the local rewiring algorithms described above. Patterns selected by design or evolution of the complex network in question would manifest themselves by $R(j) > 1$, while suppressed patterns correspond to $R(j) < 1$.

While $R(j)$ determines the magnitude of the suppression/enhancement it tells nothing about the statistical significance of the effect. This latter quantity is measured by the Z-score of the deviation:

$$Z(j) = \frac{N(j) - \overline{N_r(j)}}{\sigma_r(j)} \quad , \quad (3)$$

where $\sigma_r(j)$ is the standard deviation of $N_r(j)$ measured in a sufficiently large ensemble of randomized networks.

Alternatively the statistical significance of the difference between real and randomized networks can be quantified in terms of its P-value. The P-value is defined as the probability that the number of patterns $N_r(j)$ in a randomized network is larger or equal (or smaller or equal in case when $N(j) < \overline{N_r(j)}$) than $N(j)$. For patterns that are highly statistically significant it is often impossible to directly evaluate the P-value in a reasonable number of realizations of random networks. In this case one reports an upper bound on such a P-value given by the inverse size of the ensemble studied numerically. If one can verify that N_r is a Gaussian-distributed random variable the Z-score can be easily converted to the P-value.

In the next chapter we discuss a particular example of the topological profile of a network quantifying correlations between connectivities of its neighboring nodes. In this case $N(j)$ is given by $N(K_0, K_1)$ – the number of edges connecting nodes of connectivity K_0 to those of connectivity K_1 [8, 9];

A different topological profile was studied in [16, 17]. In their case $N(j)$ stood for the number of a particular small network motif (involving not more than 4 nodes) such as e.g. a feed-forward or a feedback triangular loop in directed networks.

Correlation profiles of protein networks.

Methods described in the previous chapter allow us to define and measure the *correlation profile* of any large complex network. The correlation profile quantifies correlations between connectivities of its neighboring nodes. We have applied these numerical tools to investigate two levels of molecular networks in yeast *Saccharomyces Cerevisiae*, which at present is perhaps the best characterized biological model organism:

1. The *protein interaction network* used in this work consists of 4475 physical interactions between 3279 yeast proteins as measured in the most comprehensive high-throughput yeast two-hybrid screen [18]. In order to better visualize the protein interaction network in Fig. 3 we plotted its small part formed by all interactions of proteins localized in the yeast nucleus [1] with each other.
2. The most general definition of the *regulatory network* includes all cases when production or degradation of one of its proteins is *directly* controlled by another. Edges of this network correspond to transcription and translational regulation, phosphorylation, allosteric modification, specific targeting for degradation of one protein by another, etc. The YPD database [1] contains 1772 (1328 positive and 444 negative) such regulations among 848 yeast proteins. To narrow down the range of possible regulatory mechanisms and make the resulting network more homogeneous we have constructed correlation profiles of the *transcription regulatory network*, which is the subset of the general regulatory network formed by all positive and negative direct transcription regulations. This network, shown in Fig. 4, consists of 1289 (1047 positive

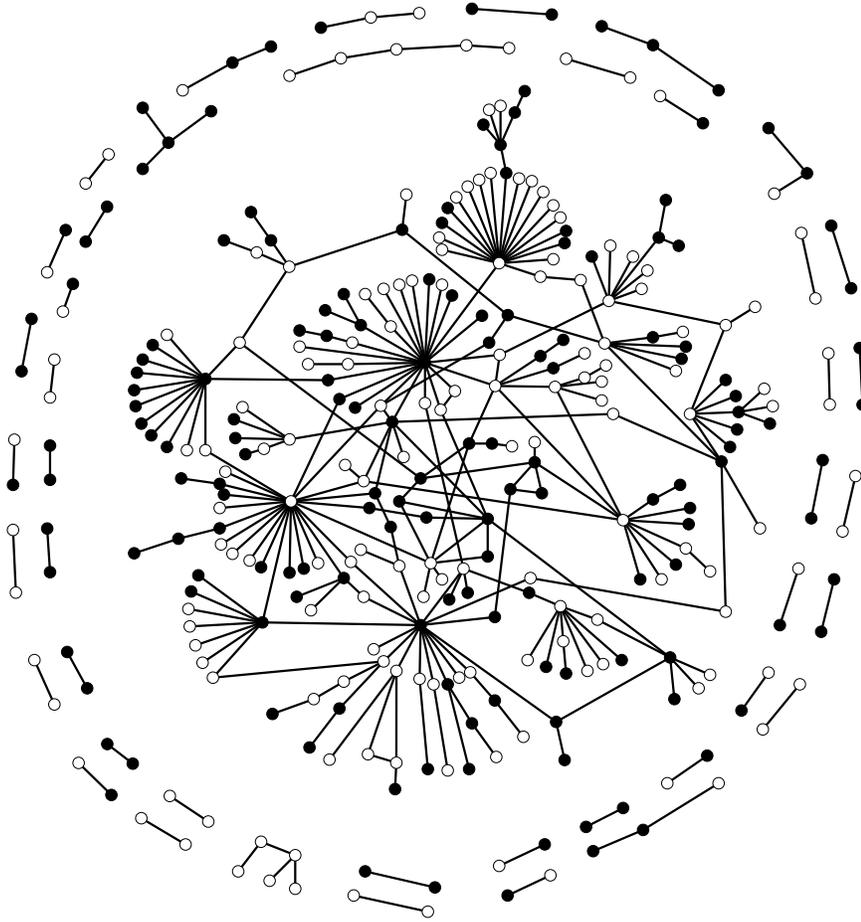


Figure 3: Network of physical interactions between nuclear proteins in yeast. Here we show the subset of protein-protein physical interactions reported in the full set of Ref. [18]. The subset consists of 318 interactions between proteins that are known to be localized in the yeast nucleus [1]. The resulting network involves 329 proteins. Note that most neighbors of highly connected proteins have rather low connectivity. This feature will be later quantified in the correlation profile of this network (Figs 5, 7). Nodes are color coded according to how essential they are for the survival of yeast cells under laboratory conditions [1]. White nodes correspond to viable and black ones to non-viable null-mutants lacking the corresponding protein.

and 242 negative) regulations by 125 transcription factors [1] within the set of 682 proteins.

While the regulatory network is naturally directed, the network of physical interactions among proteins in principle lacks directionality. Randomized versions of these two molecular networks were constructed by randomly rewiring their edges, while preventing “unphysical” multiple connections between a given pair of nodes, as described in the previous chapter. By construction this algorithm separately conserves the in- and out-connectivities of each node. Therefore, in a randomized version of the regulatory network each protein has the same numbers of regulators and regulated proteins as in the original network. Taking in consideration the bait-prey asymmetry mentioned in [8], when generating random counterpart of the interaction network we chose to separately conserve numbers of interaction partners of the bait-hybrid and the prey-hybrid of every protein. The set of MATLAB programs for both randomization algorithm and the correlation profile detection and visualization can be downloaded from [10].

The topological property of the network giving rise to its correlation profile is the number edges $N(K_0, K_1)$ connecting pairs of nodes with connectivities K_0 and K_1 . To find out if in a given complex network connectivities of interacting nodes are correlated, $N(K_0, K_1)$ should be compared to its value $N_r(K_0, K_1) \pm \Delta N_r(K_0, K_1)$ in a randomized network, generated by the edge rewiring algorithm. When normalized by the total number of edges E , $N(K_0, K_1)$ defines the joint probability distribution $P(K_0, K_1) = N(K_0, K_1)/E$ of connectivities of interacting nodes. Any correlations would manifest themselves as systematic deviations of the ratio

$$R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1) \quad (4)$$

away from 1. Statistical significance of such deviations is quantified by their Z-score

$$Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1))/\sigma_r(K_0, K_1), \quad (5)$$

where $\sigma_r(K_0, K_1) = \Delta N_r(K_0, K_1)/N$ is the standard deviation of $P_r(K_0, K_1)$ in an ensemble of randomized network.

Figs. 5 and 6 show the ratio $R(K_0, K_1)$ as measured in yeast interaction and transcription regulatory networks, respectively. In the interaction network K_0 and K_1 are numbers of neighbors of the two interacting proteins,

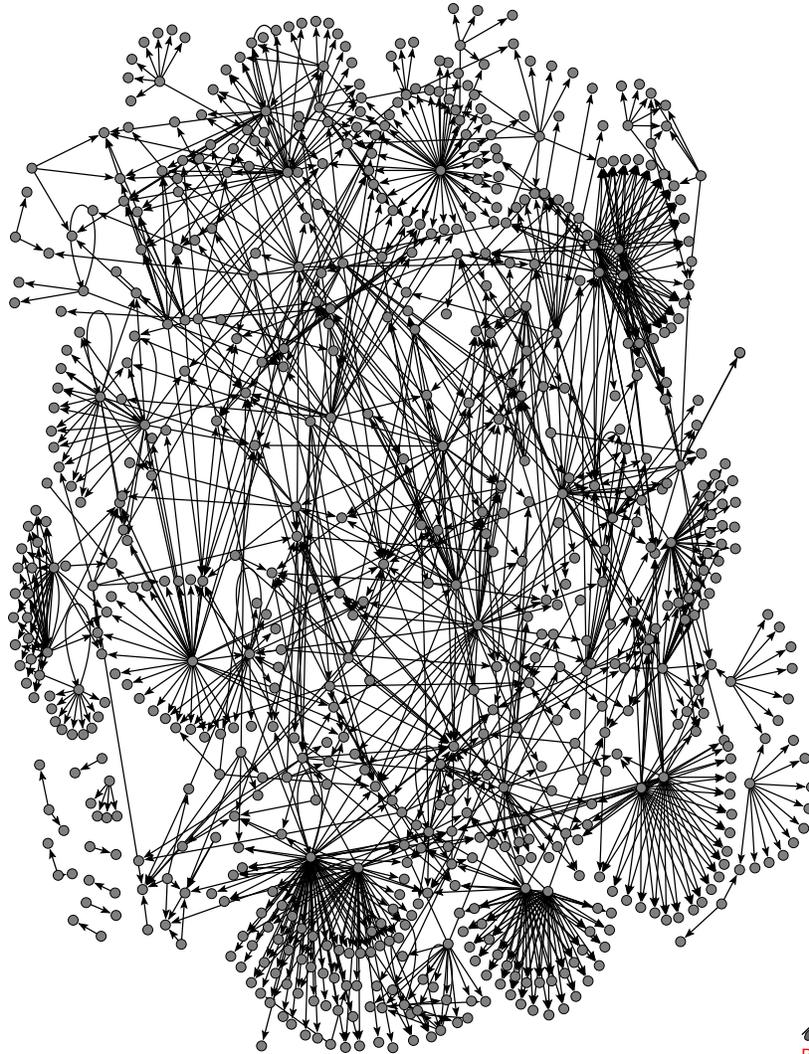


Figure 4: Transcription regulatory network in yeast. Apart from the absence of clear signs of modularity (the network has a unique giant connected component or module and only a few small disconnected modules), one notices several striking features related to hub proteins that each regulate many other proteins: 1) They tend to regulate genes with just a few regulatory inputs. As a result of this they are well separated from each other, and positioned [2] on a periphery of the network. 2) it is much more frequent for a protein to regulate many other proteins, than to be regulated by many. It is the first of these features, the separation of out- and in-hubs from each other, that is quantified with the help of the correlation profile of this network (Figs 6, 8).

while in the regulatory network K_0 is the out-connectivity of the regulatory protein and K_1 – the in-connectivity of its regulated partner. Thus by the very construction $P(K_0, K_1)$ is symmetric for the physical interaction network but not for the regulatory network. Figs. 7,8 plot the statistical significance $Z(K_0, K_1)$ of deviations visible in Figs. 5,6 correspondingly. To arrive at these Z-scores 100 randomized networks were sampled and connectivities were logarithmically binned into two bins per decade.

The combination of R - and Z -profiles reveals the regions on the $K_0 - K_1$ plane, where connections between proteins in the real network are significantly enhanced or suppressed, compared to the null model. In particular, the light region in the upper right corner of Figs. 5-8 reflects the reduced likelihood that two hubs are directly linked to each other, while dark regions in the upper left and the lower right corners of these figures reflect the tendency of hubs to associate with nodes of low connectivity. One should also note a prominent light-colored feature on the diagonal of the Fig. 5 and 7 corresponding to an enhanced affinity of proteins with between 4 and 9 physical interaction partners towards each other. This feature can be tentatively attributed to members of multi-protein complexes interacting with other proteins from the same complex. The above range of connectivities thus correspond to a typical number of direct interaction partners of a protein in a multi-protein complex. When we studied pairs of interacting proteins in this range of connectivities we found 39 of such pairs to belong to the same complex in the recent high-throughput study of yeast protein complexes [20]. This is about 4 times more than one would expect to find by pure chance alone.

When analyzing molecular networks one should consider possible sources of errors in the underlying data. Two-hybrid experiments in particular are known to contain a significant number of false positives and probably even more false negatives.

The evidence of a significant number of false negatives lies in the fact that only a small fraction of functionally plausible interactions were detected in both directions (the bait-hybrid of a protein A interacting the prey-hybrid of a protein B as well as the prey-hybrid of a protein A interacting the bait-hybrid of a protein B). It is also attested by a relatively small overlap in interactions detected in the two independent high-throughput two hybrid experiments [19, 18]. There exist a number of plausible explanations of these

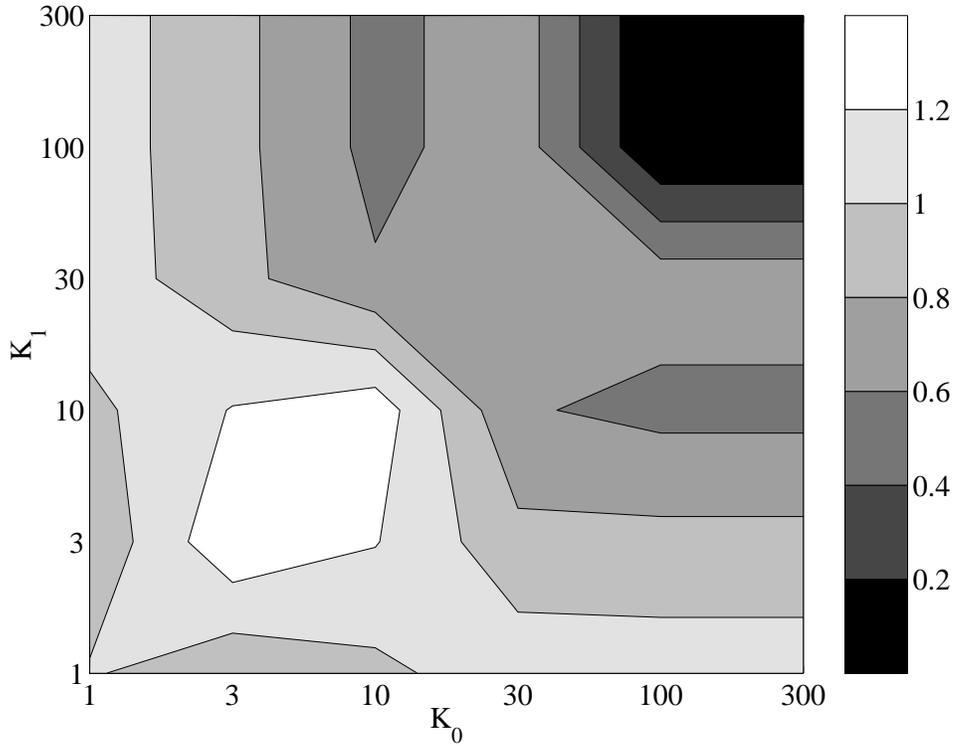


Figure 5: Correlation profile of the protein interaction network in yeast. The ratio $R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with K_0 and K_1 interaction partners correspondingly, directly interact with each other in the full set of Ref. [18], while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text. Note the logarithmic scale of both axes.

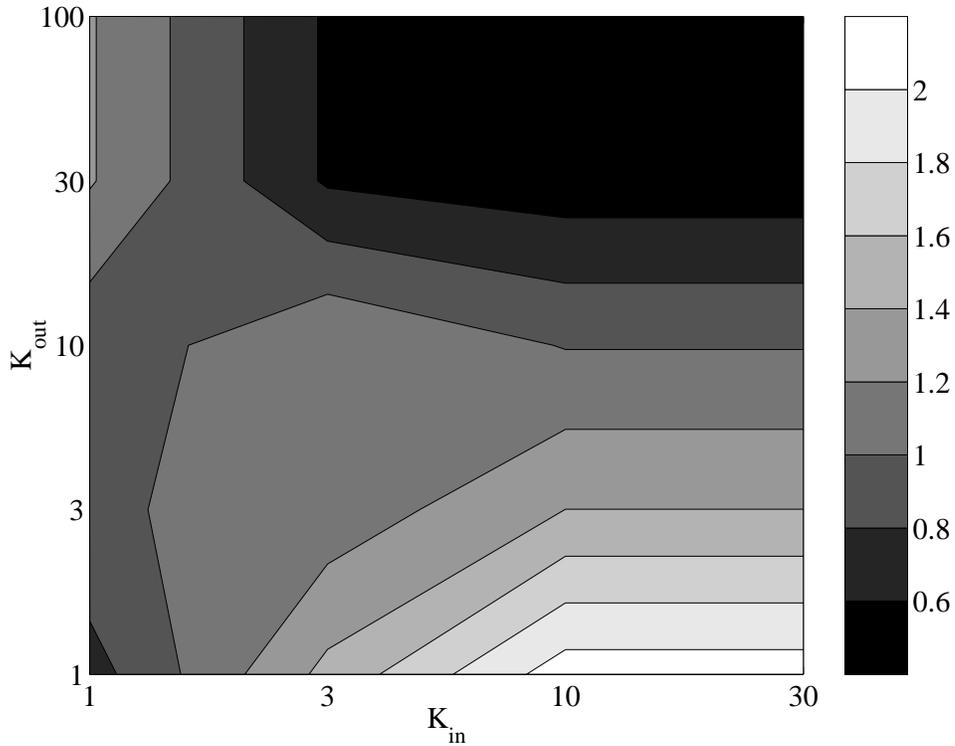


Figure 6: Correlation profile of the transcription regulatory network in yeast. The ratio $R(K_{out}, K_{in}) = P(K_{out}, K_{in})/P_r(K_{out}, K_{in})$, where $P(K_{out}, K_{in})$ is the probability that a protein node with the out-connectivity K_{out} transcriptionally regulates the protein node with the in-connectivity K_{in} in the network from the YPD database [1], while $P_r(K_{out}, K_{in})$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text. Note the logarithmic scale of both axes.

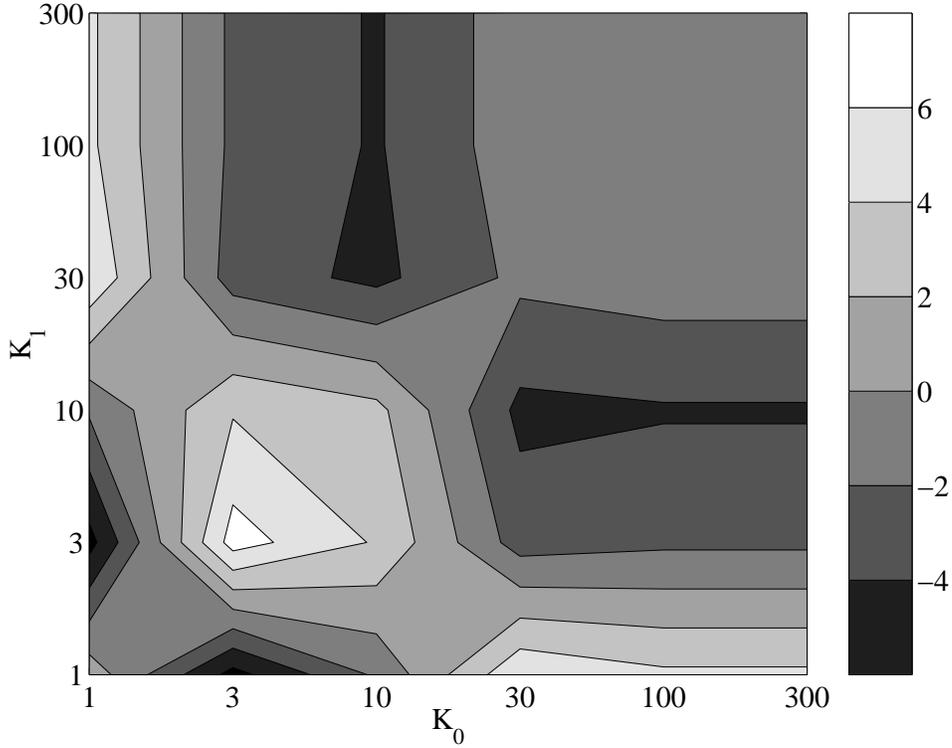


Figure 7: Statistical significance of correlations present in the protein interaction network in yeast. The Z-score of correlations $Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1)) / \sigma_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with K_0 and K_1 interaction partners correspondingly, directly interact with each other in the full set of Ref. [18], while $P_r(K_0, K_1)$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text, and $\sigma_r(K_0, K_1)$ is the standard deviation of $P_r(K_0, K_1)$ measured in 1000 realizations of a randomized network. Note the logarithmic scale of both axes.

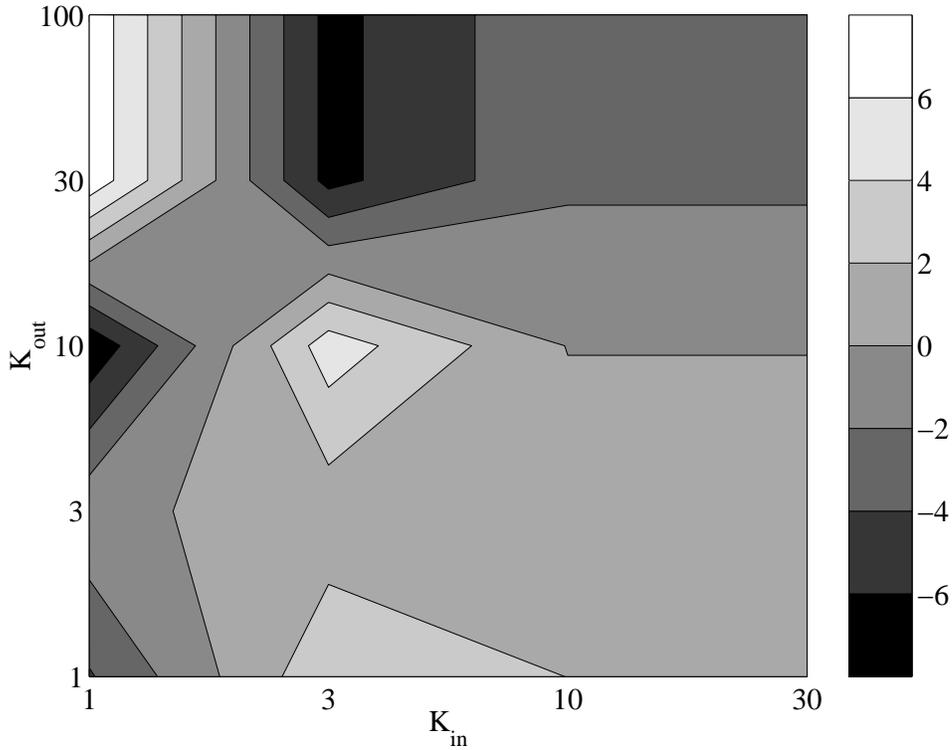


Figure 8: Statistical significance of correlations present in the transcription regulatory network in yeast. The ratio $Z(K_{out}, K_{in}) = (P(K_{out}, K_{in}) - P_r(K_{out}, K_{in})) / \sigma_r(K_{out}, K_{in})$, where $P(K_{out}, K_{in})$ is the probability that a protein node with the out-connectivity K_{out} transcriptionally regulates the protein node with the in-connectivity K_{in} in the network from the YPD database [1], while $P_r(K_{out}, K_{in})$ is the same probability in a randomized version of the same network, generated by the random rewiring algorithm described in the text, and $\sigma_r(K_{out}, K_{in})$ is the standard deviation of $P_r(K_{out}, K_{in})$ measured in 1000 realizations of a randomized network. Note the logarithmic scale of both axes.

false negatives. First of all, binding may not be observed if the conformation of the bait or prey chimeric protein blocks relevant interaction sites or if it altogether fails to fold properly. Secondly, it is not entirely clear if the number of cells in batches used in high-throughput two hybrid experiments is sufficient for any given bait-prey pair to meet in at least one cell. Finally, 391 out of potential 5671 baits in [18] were not experimentally tested because they were found to activate the transcription of the reporter gene in the absence of any prey proteins.

Several sources of false positives are also commonly mentioned in the literature:

- In one scenario spurious interactions of highly connected baits are thought to arise due to a *low-frequency* indiscriminate activation of the reporter gene in the absence of any prey proteins. Such false positives (if they exist) are easy to eliminate by using curated high-throughput datasets which contain only protein pairs that were observed, say, at least 3 times in the course of the experiment. We have shown that all qualitative features of the correlation profile of the protein interaction network reported above remain unchanged when one uses such curated datasets [21].
- In another scenario the interaction between proteins is real but it never happens in the course of the normal life cycle of the cell due to spatial or temporal separation of participating proteins. However, it is hard to believe that such non-functional interactions would be preserved for a long time in the course of evolution. Hence, it is dubious that such false-positives would be ubiquitous.
- In yet another scenario an indirect physical interaction is mediated by one or more unknown proteins localized in the yeast nucleus. However, since in two-hybrid experiments bait and prey proteins are typically highly overexpressed, it is only very abundant intermediate proteins that can give rise to an indirect binding. The relative insignificance of indirect bindings is attested by a relatively small number of triangles (178 vs $\propto 100$ in a randomized version) in the protein interaction network. Indeed, an indirect interaction of a protein A with a protein B effectively closes the triangle of direct interactions A-C and C-B with an intermediate protein C.

1 Discussion: What it may all mean?

The large-scale organization of molecular networks deduced from correlation profiles of protein interaction and transcription regulatory networks in yeast is consistent with compartmentalization and modularity characteristic of many cellular processes [22]. Indeed, the suppression of connections between highly-connected proteins (hubs) suggests the picture of semi-independent modules centered around or regulated by individual hubs. On the other hand, the very fact that these molecular networks do not separate into many isolated components but are dominated by one “giant component” suggests that this tendency towards modularity is not taken to its logical end. The observed patterns can in fact be characterized as “soft modularity”, where interactions between individual modules are suppressed but not completely eliminated. Thus on sufficiently large scale molecular networks exhibit system-wide properties making their behavior different from that of a set of mutually independent modules. Two recent empirical observations hint at global interrelations in the overall connectivity pattern of molecular networks:

1. Elena and Lenski [23] studied the cooperativity of regulation in *E.coli* by comparing changes of the cell cycle length in single-gene null mutants with those in double null mutants. They concluded that about 30% of gene pairs exhibited more than additive effects on cell cycle length, and thus at least 30% of protein pairs are functionally interconnected. Such level of cooperativity would be impossible in a regulatory network consisting of a large number of independent modules.
2. C.K. Stover et al. [24] found that the number of transcription factors (N_{tr}) in procaryotic organisms grows as a *square* of the number of genes (N): $N_{tr} \propto N^2$. Hence, each additional gene (or a module of functionally related genes) appears to be regulated with respect to all other genes present in the genome. This indicates an overall regulation pattern that on sufficiently large scale is neither modular, nor hierarchic. The equation $N_{tr}/N = \langle K_{in} \rangle / \langle K_{out} \rangle$ relates the fraction of transcription factors in the genome to the average in- and out-connectivities of the transcription regulatory network. On the network level the growth of $N_{tr}/N \propto N$ with N is most naturally achieved by an increase in complexity of regulation of individual genes: $\langle K_{in} \rangle$. Thus regulatory

networks inevitably become more and more interconnected in more and more complex organisms.

A further implication of the deficit of connections between highly connected proteins (Figs. 5, 6) is in the suppression of propagation of deleterious perturbations over the network. It is reasonable to assume that certain perturbations such as e.g. a significant change in the concentration of a given protein (including it vanishing altogether in a null-mutant cell) with a certain probability can affect its first, second, and sometimes even more distant neighbors in the corresponding network. While the number of immediate neighbors of a node is by definition equal to its own connectivity K_0 , the average number of its second neighbors is bound from above by $K_0 \langle (K_1 - 1) \rangle_{K_0}$ and thus depends on the correlation profile of the network. Since highly connected nodes serve as powerful amplifiers for the propagation of deleterious perturbations it is especially important to suppress this propagation beyond their immediate neighbors. It was argued that scale-free networks in general are very vulnerable to cascading failures started at individual hubs [25, 26]. The deficit of edges directly connecting hubs to each other reduces the branching ratio around these nodes and thus provides a certain degree of protection against such accidents.

To summarize the above discussion, it is feasible that molecular networks operating in living cells have organized themselves in an interaction pattern that is both robust and specific. Topologically the specificity of different functional modules is enhanced by limiting interactions between hubs and suppressing the average connectivity of their neighbors. On a larger scale there is evidence for interconnections between these modules, although the principles of such global organization of living cells remain unclear from the present day data and analysis tools.

The main goal of the present review was to introduce a number of statistical tools necessary for analyzing topological patterns and correlations present in biological networks. These tools allowed us to identify the set of distinctive topological features of several protein networks, which may help to better understand possible mechanisms of their function and evolution. The advantage of our approach lies also in its iterative nature in which the understanding of more and more complex topological patterns gradually builds up on the analysis of the lower level features.

References

- [1] As reported in the YPD database: Costanzo, M. C. *et al.* (2001) *Nucleic Acids Research* **29**, 75-79.
- [2] Batagelj, V. and Mrvar, A. (1998) *Connections* **21** 2, 47-57.
- [3] Barabasi, A.-L. , Albert, R. (1999) *Science* **286**, 509–512.
- [4] Jeong, H., Tombor, B. , Albert, R., Oltvai, Z. N. , Barabasi, A.-L. (2000) *Nature* **407**, 651–654.
- [5] Jeong, H. , Mason, S., Barabasi, A.-L., Oltvai, Z.N. (2001) *Nature* **411**, 41–42.
- [6] Eguiluz, V. M., Cecchi, G. Chialvo, D. R. et al. Preprint at arXiv.org e-Print archive available at <http://arxiv.org/abs/cond-mat/0309092> (2003).
- [7] Newman, M. E. J., Strogatz, S. H. and Watts, D. J. (2001) *Phys. Rev. E*, **64**, 026118, pp.1-17.
- [8] Maslov, S. and Sneppen, K. (2002) *Science* **296**, 910-913.
- [9] Maslov, S., Sneppen, K. and Zaliznyak, A. (2002) Preprint at arXiv.org e-Print archive available at <http://arxiv.org/abs/cond-mat//0205379>; (2003) *Physica A in press*.
- [10] The set of MATLAB programs can be downloaded at <http://www.cmth.bnl.gov/maslov/matlab.htm>
- [11] Erdős, P. and Rényi, A. (1960) *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 1760.
- [12] Molloy, M. and Reed, B. (1995) *Random Struct. Algorithms* **6**, 161;
Molloy, M. and Reed, B. (1998) *Combinatorics, Probab. Comput.* **7**, 295.
- [13] Early studies of these algorithms in the context of matrices were reported in: Gale, D. (1957) *Pacific J. Math.* **7**, 1073-1082 ; Ryser, H.J. (1964) in *Recent Advances in Matrix Theory*, pp. 103-124, Univ. of Wisconsin Press, Madison, WI. For more recent references including applications

to graphs see e.g.: Kannan, R., Tetali, P., Vempala, S. (1999) Random Structures and Algorithms **14**, 293-308.

- [14] Metropolis, N., *et al.*, (1953) J. Chem. Phys. **21**, 1087.
- [15] Watts, D. and Strogatz, S. (1998) Nature **293**, 400-403.
- [16] Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002) Nature Genetics, **31**(1):64-68.
- [17] Milo, R., *et al.* (2002) Science **298**, 824-827.
- [18] Ito, T., *et al.*, (2001) Proc. Natl. Acad. Sci. USA **98**, 4569-4574.
- [19] Uetz, P., *et al.*, (2000) Nature **403**, 623-627.
- [20] Gavin, A.-C., *et al.*, (2002) Nature **415**, 141-147.
- [21] Maslov, S. and Sneppen, K. (2002) FEBS Letters **530**, 255-256.
- [22] Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999) Nature **402** (6761 Suppl), C47-52.
- [23] Elena, S.F. and Lenski, R.E. (1999) Nature **390**, 395-398.
- [24] Stover, C.K., *et al.*, (2000) Nature **406**, 959-398.
- [25] Albert, R., Jeong, H., and Barabasi, A.-L. (2000) Nature **406**, 378-382.
- [26] Vogelstein, B., Lane, D., and Levine, A.J. (2000) Nature **408**, 307-310.