BNL-71827-2003-BC

# LARGE-SCALE TOPOLOGICAL PROPERTIES OF MOLECULAR 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 necessary 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 expresses herein do not necessarily state to reflect those of the United States Government or any agency thereof.

# Large-scale topological properties of molecular networks

Sergei Maslov[1], Kim Sneppen[2]

Abstract:

Bio-molecular networks lack the top-down design. Instead, selective forces of biological evolution shape them from raw material provided by random events such as gene duplications and single gene mutations. As a result individual connections in these networks 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, ans/or its fundamental design principles and limitations?

Here we introduce a general method allowing one to construct a random null-model version of a given network while preserving the desired set of its low-level topological features, such as, e.g., the number of neighbors of individual nodes, the average level of modularity, preferential connections between particular groups of nodes, etc. Such a null-model network can then be used to detect and quantify the non-random topological patterns present in large networks.

In particular, we measured correlations between degrees of interacting nodes in protein interaction and regulatory networks in yeast. It was found that in both these networks, links between highly connected proteins are systematically suppressed. This effect decreases the likelihood of cross-talk between different functional modules of the cell, and increases the overall robustness of a network by localizing effects of deleterious perturbations. It also teaches us about the overall computational architecture of such networks and points at the origin of large differences in the number of neighbors of individual nodes.

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

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

# 1 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 the physical interaction with its 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 exert control on 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.

- On yet higher level individual cells in 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 intercellular 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.

In this review we concentrate on large-scale topological properties of complex biological networks operating on the levels of physical protein-protein interactions and transcriptional regulation.

# 2 Topological properties of protein networks

## 2.1 Single-node topological properties

An interesting property of many biological networks that was recently brought to attention of the scientific community [1, 2, 3] is an extremely broad distribution of nodes' degrees (often called connectivities in the network literature) 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 degree of the most connected hub in such a network is typically several orders of magnitude larger than the average degree in the network. Often the number of nodes $N(K)$ with a given degree $K$ can be approximated by a scale-free power law form $N(K) \sim K^{-\gamma}$ in which case the network is referred to as scale-free [1].

In this review we concentrate on large-scale properties of physical interaction and regulatory protein networks. In Fig 1 we show the presently known [4] set of transcriptional regulations in a procaryotic bacterium *Escherichia coli*. For comparison, Fig. 2 shows the presently known [5] transcriptional regulations in a simple single-cell eucaryote, *Saccharomyces cerevisiae* (baker's yeast).

Both yeast and *E.coli* regulatory networks are characterized by the above mentioned broad distribution of out-degrees $K_{out}$ of its protein-nodes defined as the number of directed arrows emanating from individual regulatory proteins. Clearly visible in Figs. 1,2 are the hub regulatory proteins that control the expression level of an unusually large number other proteins. For example, in the *E.coli* network one can see an extremely highly connected node in the lower half of Fig. 1. It is the CAP protein that senses the glucose level, and in response to it orchestrates a cooperative action of a large battery of other proteins related to its utilization.

By comparing Figs. 1, and 2 one gets an impression that the apparent growth in complexity of the transcription regulatory network from procaryotes to eucaryotes is achieved mostly by the virtue of an increase in the typical number of regulatory inputs of a protein (in-degree) $K_{in}$.

To quantify this further in Fig. 3A we compare distributions of nodes' in-degrees in transcriptional regulatory networks of yeast (diamonds, dashed-line) and E. coli (circles, solid-line). One can clearly see that the distribution of in-degrees in yeast is significantly broader than that in the E. coli. Indeed,

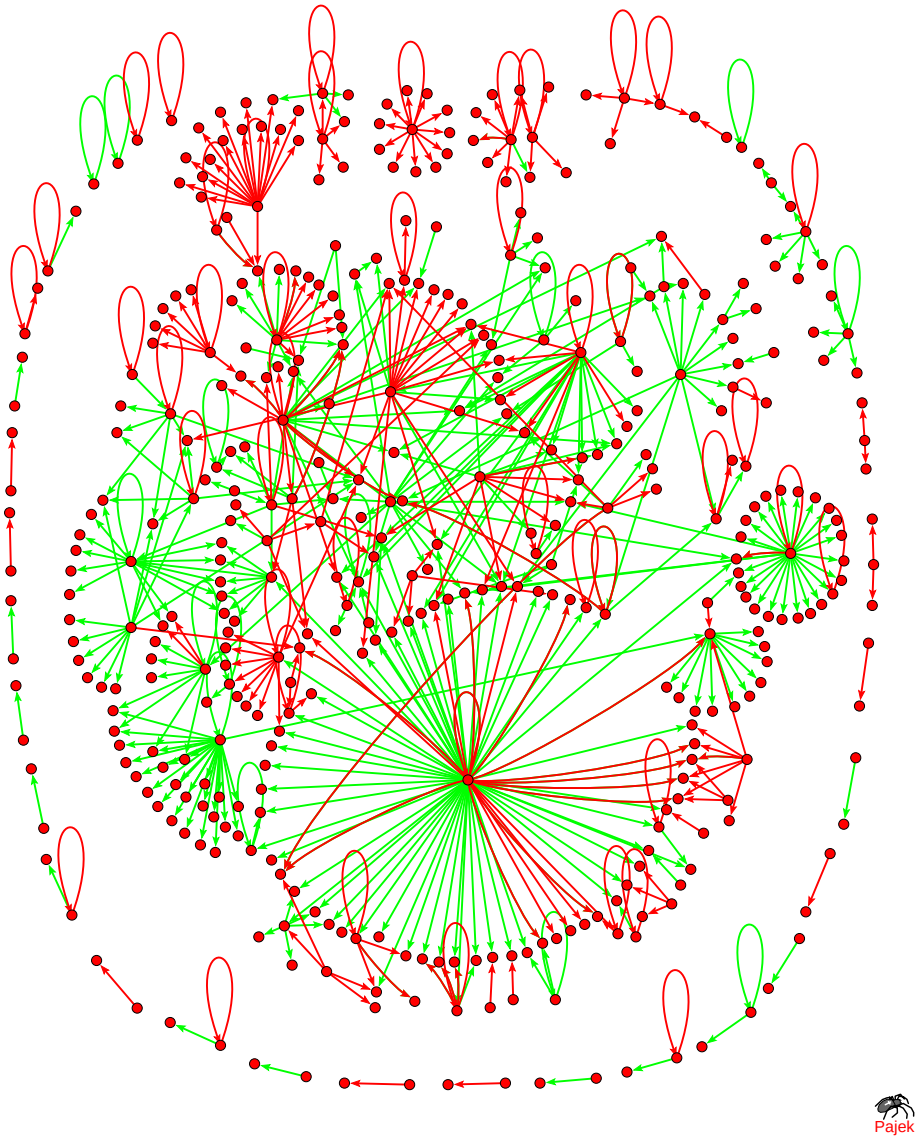


Figure 1: Presently known [4] transcriptional regulations in *E. coli*. Green and red arrows denote positive and negative regulations correspondingly. Nodes in this network represent operons (groups of genes transcribed onto a single mRNA) and arrows (edges) – direct transcriptional regulation of a downstream operon by a transcription factor encoded in the upstream operon. This network consists of 606 regulations of 424 operons by transcription factors contained in 116 different operons.

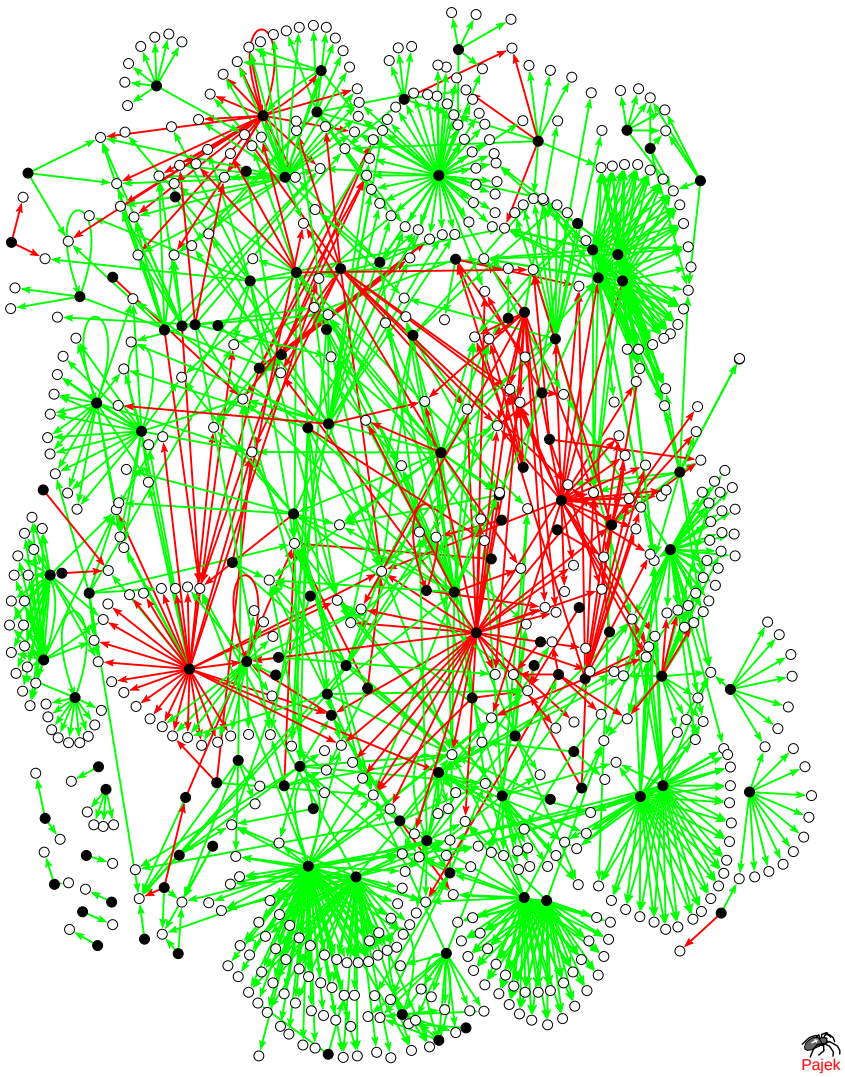


Figure 2: Presently known [5] transcriptional regulations in baker's yeast *S. cerevisiae*. This network consists of 1289 regulations of 682 proteins by 125 transcription factors. Green and red arrows denote positive and negative regulations correspondingly. Vertices corresponding to transcription factors are filled while those of remaining proteins are left empty. Apart from the absence of clear signs of modularity (the network has a unique giant connected component or module and only a few small 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 form each other, and positioned on a periphery of the network. This will be later quantified in the correlation profile of this network (Figs 7, 9). 2) It is much more frequent for a protein to regulate many other proteins, than to be regulated by many.

while in the E. coli $K_{in}$ has an exponential distribution ranging only between 0 and 6, in yeast its range is already between 0 and 15 and the tails of the distribution start to significantly deviate from the exponential functional form.

The above observations are in agreement with two recent empirical studies: C.K. Stover et al. [6] found that the number of transcription factors ($N_{tr}$) in procaryotic organisms grows as a *square* of the number of genes ($N$): $N_{tr} \propto N^2$. Very recently E. van Nimwegen [7] has extended this result to eucaryotes where he also observed a superlinear scaling $N_{tr} \propto N^{1.26}$. The exact equation

$$N_{tr}/N = \langle K_{in} \rangle / \langle K_{out} \rangle \tag{1}$$

relates the fraction of transcription factors in the genome of an organism to the average in- and out-degrees of its transcription regulatory network. Thus a direct consequence of the growth of the ratio $N_{tr}/N$ with $N$ is the increase in complexity of regulation of individual genes: $\langle K_{in} \rangle$.

The distribution of $K_{out}$ shown in Fig. 3B appears to be about equally broad in E. coli and yeast. It ranges between 1 and about 70 regulations in both networks. The power-law fit $N(K_{out}) \sim K_{out}^{-\gamma}$ gives $\gamma \approx 2$ in E-coli, while in yeast the distribution seems to have an initial slope characterized by $\gamma \approx 1$ followed by a sharper decay for $K_{out} > 30$. However, due to a limited range and an incomplete and possibly anthropomorphically biased nature of the data (databases of research articles) one should not take these fits too seriously: at the very least they indicate an unusually broad distribution of degrees in both networks.

Comparison of the Fig. 3 A and B also shows that in both organisms the in-degree distribution is much more narrow than that of the out-degree. That is a simple consequence of the fact that regulatory proteins (those with a non-zero $K_{out}$) constitute just a small fraction of all proteins in the cell.

Apart from transcriptional regulatory networks, metabolic networks [2], and protein-protein physical interaction networks [3] are characterized by a very broad distribution in the number of neighbors of their individual nodes. A small part of such physical interaction network in baker's yeast is visualized in Fig. 4.

One aspect of a broad distribution of node degrees in protein interaction and regulatory networks, is the possibility of amplification and exponential

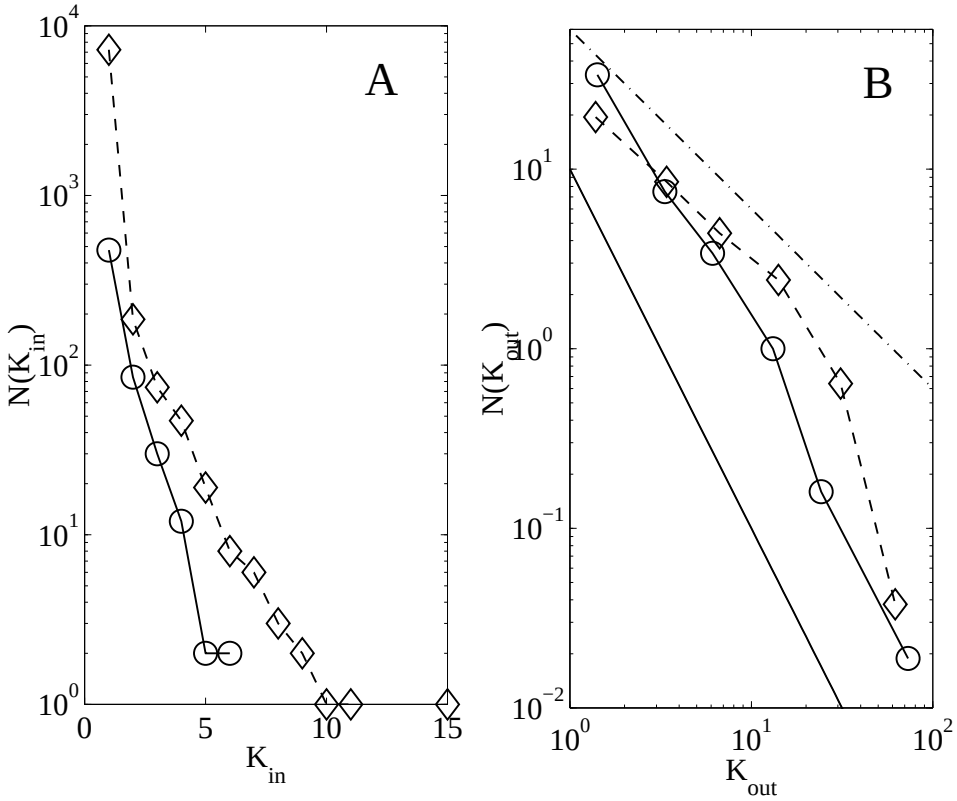


Figure 3: A. The histogram $N(K_{in})$ of nodes' in-degrees $K_{in}$ in transcription regulatory networks of yeast (diamonds, dashed line), and E. coli (circles, solid line). This histogram in yeast is noticeably broader and has fatter non-exponential tails than that in the E. Coli.
B. The histogram $N(K_{out})$ of nodes' out-degrees $K_{out}$ in transcription regulatory network in yeast (diamonds, dashed line), and E. coli (circles, solid line). Overall, these histograms are rather similar between these two very different organisms. Straight lines are power law fits with the slope $-2$ (solid) and $-1$ (dot-dashed). To improve the statistics all histograms in this panel were logarithmically binned into 5 bins per decade.

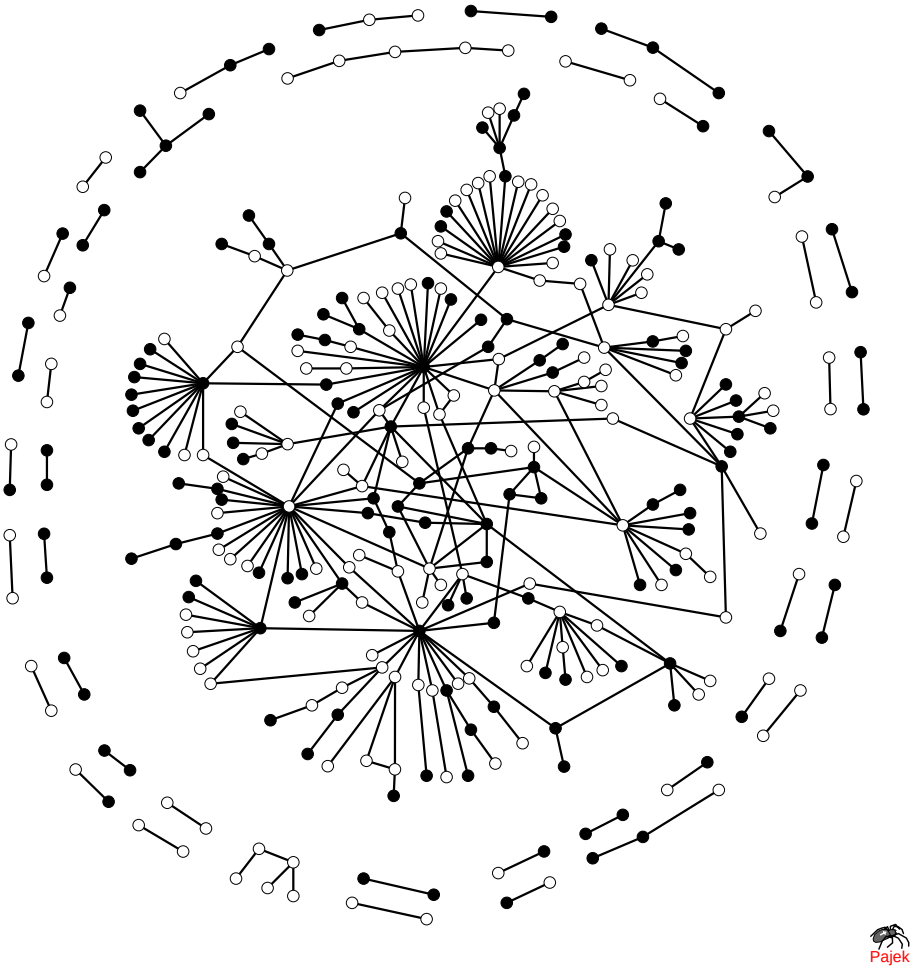


Figure 4: Network of physical interactions between nuclear proteins in yeast. Here we show the subset of protein-protein physical interactions reported in the full set of Ref. [8] consisting of 318 interactions between proteins that are known to be localized in the yeast nucleus [5]. The resulting network 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 6, 8). Nodes are color coded according to how essential they are for the survival of yeast cells under laboratory conditions [5]. White nodes correspond to viable and black ones to non-viable null-mutants lacking the corresponding protein.

spread of signals propagating in the network. The upper bound of the one step amplification of some biochemical signal propagating in a *directed* network is given by

$$\mathcal{A}^{(dir)} = \frac{\langle K_{in} K_{out} \rangle}{\langle K_{in} \rangle} \quad . \tag{2}$$

This amplification factor $\mathcal{A}^{(dir)}$ measures the average number of neighbors to which the signal can be potentially broadcasted in one propagation step. The above formula, derived by Newman in Ref. [9], follows from the observation that a signal enters a given node with a probability proportional to its in-degree $K_{in}$, and leaves along any of its $K_{out}$ outgoing links. For $\mathcal{A}^{(dir)} \leq 1$ the any given signal eventually dies out and hence affects only a small fraction of nodes in the network. On the other hand, for $\mathcal{A}^{(dir)} > 1$ signals propagating in the network might be exponentially amplified, and thus each of them could influence (and possibly interfere with) other signals over the entire network.

The degree $K$ in *undirected* networks cannot be decomposed into in- and out- components. Hence the upper bound on amplification of signals is given by the amplification factor $\mathcal{A}^{(undir)}$ [9]:

$$\mathcal{A}^{(undir)} = \frac{\langle K(K-1) \rangle}{\langle K \rangle} \quad . \tag{3}$$

In the above equation we take into account the fact that the signal cannot reach new nodes along the edge by which it came to a given node. Hence the use of $K-1$ in the enumerator. The amplification factor $\mathcal{A}^{(undir)}$ in scale-free networks with $\gamma < 3$ is very large and sensitive to the degrees of the highest connected hub-nodes. Here the borderline case $\mathcal{A}^{(undir)} = 1$ also separates two different regimes. For $\mathcal{A}^{(undir)} < 1$ the network breaks apart into many components isolated from each other, while for $\mathcal{A}^{(undir)} \gg 1$ it consists of a unique "giant" component, containing the majority of all nodes, and a few small disconnected components.

The direct calculation of the directed amplification ratio $\mathcal{A}^{(dir)}$ in the transcription regulatory network gives $\mathcal{A}_{ec}^{(dir)} = 1.08$ in the E. coli and $\mathcal{A}_{sc}^{(dir)} = 0.58$ in yeast. Hence as directed networks they are both below or approximately at (in E. coli) the critical point $\mathcal{A}_c = 1$. Therefore, signals propagating in these networks cannot exponentially amplify, which limits the extent of cross-talk between them. However if both these regulatory networks are treated as undirected (i.e. one temporarily forgets about the arrows on their edges) one gets significantly overcritical amplification ratios $\mathcal{A}^{(undir)} >> 1$:

$\mathcal{A}_{ec}^{(undir)} = 10.5$ in the E.coli and $\mathcal{A}_{sc}^{(undir)} = 13.4$ in yeast. This explains why the majority of nodes in Figs 1,2 belong to the largest connected component, and why the size of disconnected components is so small. Apparently the cross-talk presents much bigger poetntial problem in the network of physical interactions between yeast proteins (Fig. 4), where $\mathcal{A}_{PPI}^{(undir)} = 26.3$. In the last chapter of this review we would return to the question of cross-talk and demonstrate how higher-level topological properties detected in both physical and regulatory networks in yeast [10] help to reduce such undesirable interference between signals.

## 2.2 Local rewiring algorithm: constructing a randomized null-model network

The set of degrees 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 chapter 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, and which arose due to the network growth, evolution, ans/or its fundamental design principles and limitations?

To this end we first construct a proper randomized version (null model) of a given network. As was pointed out in the general context of complex scale-free networks [9], a broad distribution of degrees indicates that the degree itself is an important individual characteristic of a node and as such it should be preserved in the randomized null-model network [10]. In addition to degrees one may choose to preserve some other low-level topological properties of the network in question [11]. Any measurable topological quan-

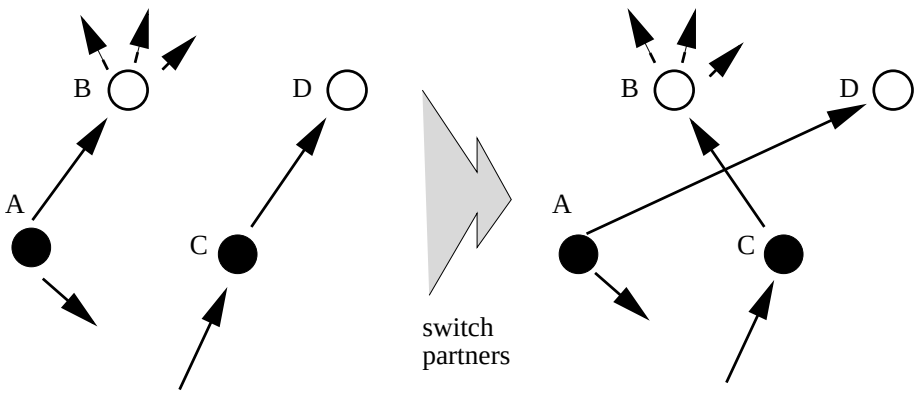


Figure 5: One step of the random local rewiring algorithm. A pair of edges A→B and C→D is randomly selected. The two edges are then rewired in such a way that A becomes connected to D, while C to B, provided that none of these new edges already exist in the network, in which case the rewiring step is aborted and a new pair of edges is selected. An independent random network is obtained when the above local switch move is performed 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- degrees of each individual node.

tity, such as e.g. the total number of edges connecting pairs of nodes with given degrees, 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 [10, 11, 12, 13].

An algorithm giving rise to a random network with the same set of individual node degrees as in a given complex network was proposed in [14, 10]. It consists of multiple repetitions of the following simple switch move (elementary rewiring step) illustrated in Fig. 5:

**Randomly select a pair of edges A→B and C→D and rewire them in such a way that A becomes connected to D, while C to B.** To prevents the appearance of multiple edges connecting the same pair of nodes, the rewiring step is aborted and a new pair of edges is selected if one or two

of the new edges already exist in the network. A repeated application of the above rewiring step 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 [15].

Sometimes it is desirable that the null-model random network in addition to nodes' degrees conserves some other topological quantity of the real network. In this case one could supplement [11] the random rewiring algorithm described above with the Metropolis acceptance/rejection criterion [16] of a switch move.

For the sake of concreteness let's assume that one wants to generate a random network with the same set of nodes' degrees and the same number $N_\triangle$ of triangles as the real undirected network [11]. Indeed, the number of triangles in a network is related to its "clustering coefficient" routinely used as a measure of its modularity [17]. Hence, by conserving $N_\triangle$ one generates a null-model with the same average level of modularity as the original complex network.

The Metropolis version [11] of the random rewiring algorithm uses an artificial energy function $H$ that favors the number of triangles in a random network $N_\triangle^{(r)}$ to be as close as possible to its value $N_\triangle$ in the real network:

$$H \;=\; \frac{\left(N_\triangle^{(r)} \;-\; N_\triangle\right)^2}{N_\triangle} \qquad . \tag{4}$$

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 $\Delta H$ 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 a local (often suboptimal or non-representative) energy minimum. In order to get a random network with $N_\triangle^{(r)}$ sufficiently close to $N_\triangle$ 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 $N_\triangle^{(r)} = N_\triangle$.

## 2.3 Multi-node properties of protein networks: correlation profile.

The *correlation profile* of any large complex network quantifies correlations between degrees of its neighboring nodes. We have calculated correlation profiles of:

1. The *protein interaction network* consisting of 4475 physical interactions between 3279 yeast proteins as measured in the most comprehensive high-throughput yeast two-hybrid screen [8]. A subset of this network is shown in Fig. 4

2. The transcriptional regulatory network in yeast (Fig. 2), consists of 1289 (1047 positive and 242 negative) regulations by 125 transcription factors [5] 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-degrees 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 [10], 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 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 degrees $K_0$ and $K_1$. To find out if in a given complex network the degrees 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 degrees of interacting nodes. Any correlations would manifest themselves as systematic deviations of the ratio

$$R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1) \tag{5}$$

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

$$Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1))/\sigma_r(K_0, K_1) \ , \tag{6}$$

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 networks.

Figs. 6 and 7 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, while in the regulatory network $K_0$ is the out-degree of the regulatory protein and $K_1$ – the in-degree of its regulated partner. Thus by its very construction $P(K_0, K_1)$ is symmetric for the physical interaction network but not for the regulatory network. Figs. 8,9 plot the statistical significance $Z(K_0, K_1)$ of deviations visible in Figs. 6,7 correspondingly. To arrive at these Z-scores 1000 randomized networks were sampled and degrees 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 blue/green region in the upper right corner of Figs. 6-9 reflects the reduced likelihood that two hubs are directly linked to each other, while red regions in the upper left and the lower right corners of these figures reflect the tendency of hubs to associate with nodes of low degree. One should also note a prominent feature on the diagonal of the Fig. 6 and 8 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 degrees 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 degrees we found 39 of such pairs to belong to the same complex in the recent high-throughput study of yeast protein complexes [18]. This is about 4 times more than one would expect to find by pure chance alone.

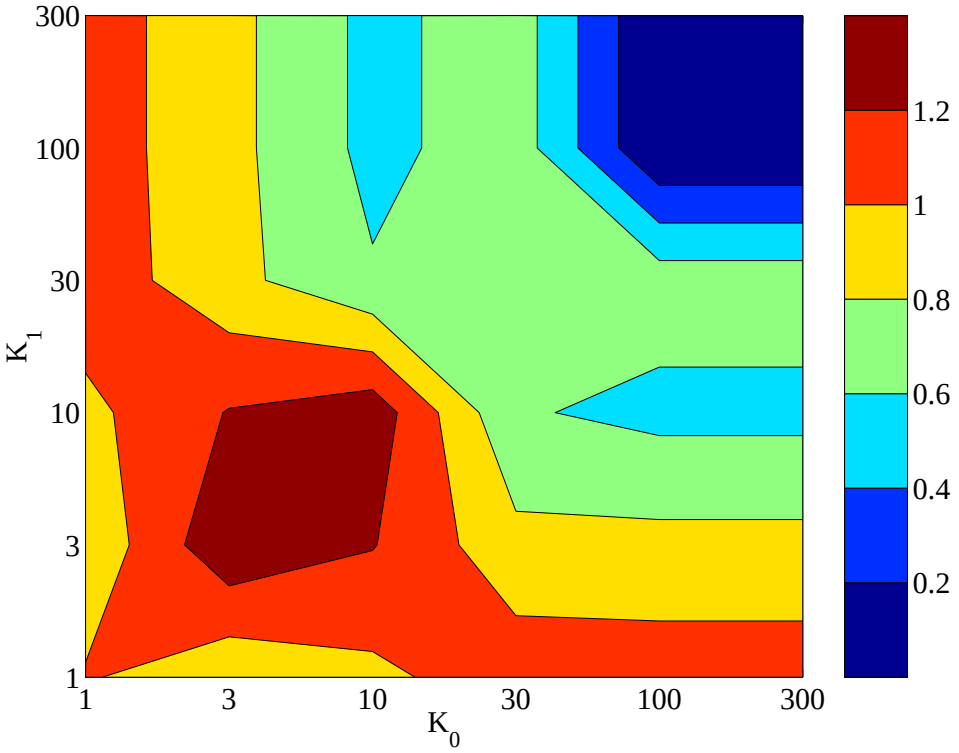


Figure 6: Correlation profile of the protein interaction network in yeast. The ratio $R(K_0, K_1) = P(K_0, K_1)/P_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with $K_0$ and $K_1$ interaction partners correspondingly, directly interact with each other in the full set of Ref. [8], 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.

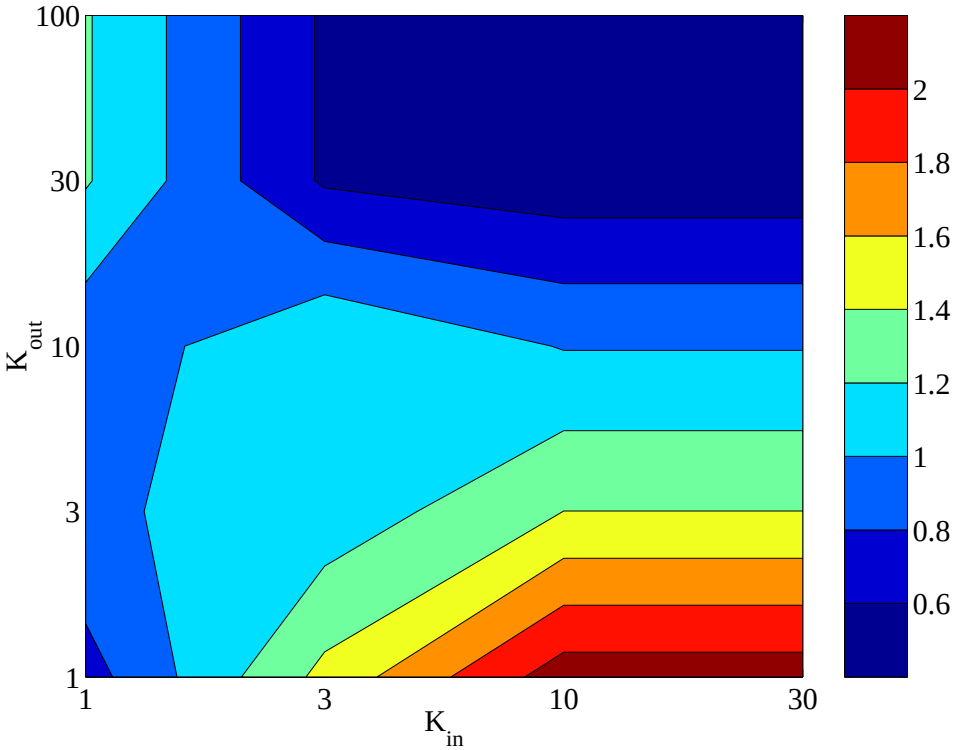


Figure 7: Correlation profile of the transcription regulatory network in yeast. The ratio $R(K_{out}, K_{in}) = P(K_{out}, K_{in})/P_r(K_{out}, K_{in})$, where $P(K_{out}, K_{in})$ is the probability that a protein node with the out-degree $K_{out}$ transcriptionally regulates the protein node with the in-degree $K_{in}$ in the transcription regulatory network obtained from the YPD database [5] (Fig. 2), 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.

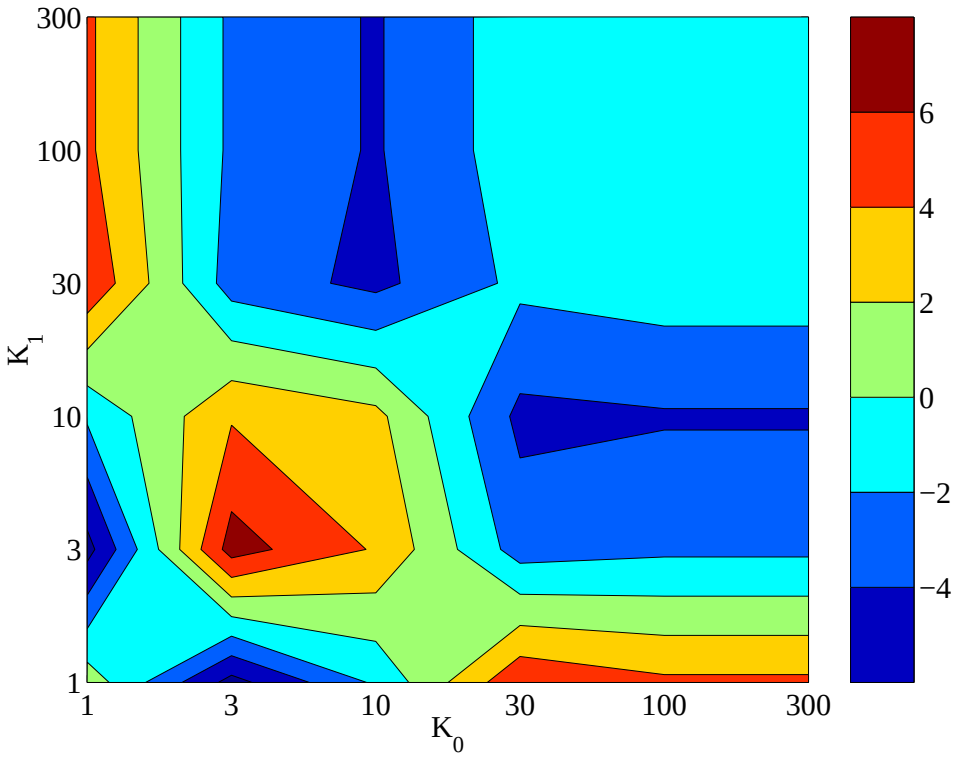


Figure 8: Statistical significance of correlations present in the protein interaction network in yeast. The Z-score of correlations $Z(K_0, K_1) = (P(K_0, K_1) - P_r(K_0, K_1))/\sigma_r(K_0, K_1)$, where $P(K_0, K_1)$ is the probability that a pair of proteins with $K_0$ and $K_1$ interaction partners correspondingly, directly interact with each other in the full set of Ref. [8], 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.

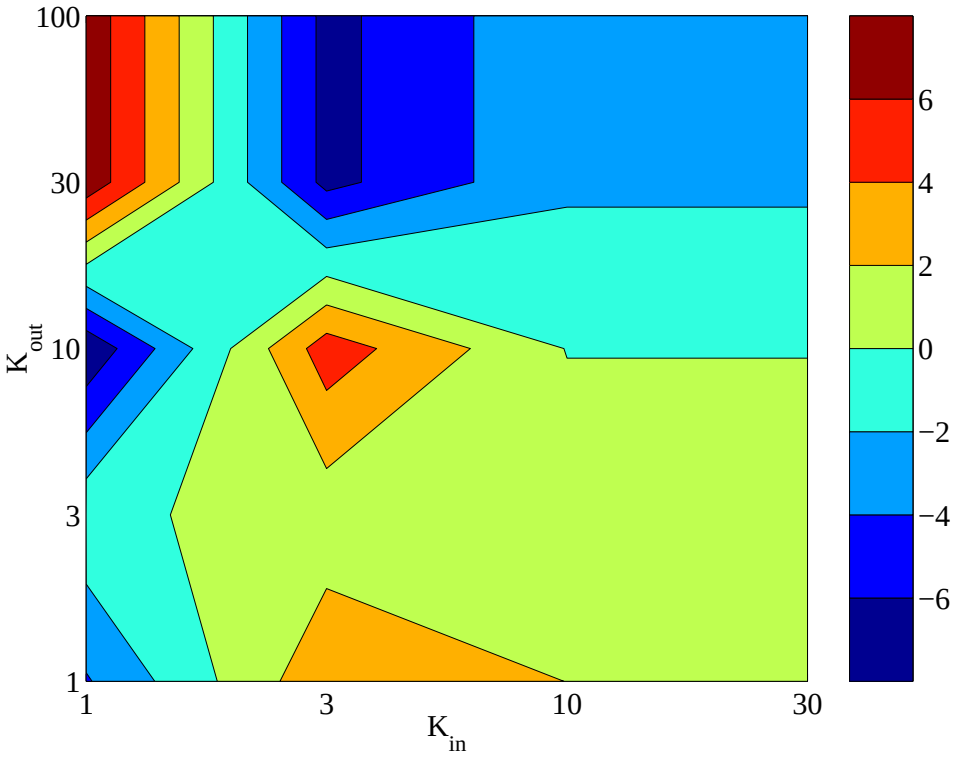


Figure 9: Statistical significance of correlations present in the transcription regulatory network in yeast. The ratio $Z(K_{out}, K_{in}) = (P(K_{out}, K_{in}) - P_r(K_{out}, K_{in}))/\sigma_r(K_{out}, K_{in}))$, where $P(K_{out}, K_{in})$ is the probability that a protein node with the out-degree $K_{out}$ transcriptionally regulates the protein node with the in-degree $K_{in}$ in the network from the YPD database [5], 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.

## 2.4 Robustness of the correlation profile with respect to potential errors in the data

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 of 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, 8]. There exist a number of plausible explanations of these 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 [8] 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 [20].

- 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.

# 3 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 [21]. 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.

A further implication of the deficit of connections between highly connected proteins (Figs. 6, 7) 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 ceratin 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 degree $K_0$, 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 [22, 23]. 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.

Finally, we would like to mention that the tendency of highly connected proteins to be positioned at the periphery of signalling and regulatory networks teaches us something about the overall computational architecture of such networks and origins of their broad degree distributions. Indeed, the peripheral position of hubs indicates that they presumably execute collective orders of other more "computationally-involved" regulators, rather than performing computations and making decision on their own. This principle is nicely illustrated in the lambda-phage regulatory network (see Fig. 10), where the decision making/computation is done by CI , CII, and Cro proteins of low-to-intermediate degree, whereas their orders are executed through the N, and LexA hub-proteins. Broad degree distributions observed in molecular networks presumably reflect the widely different needs associated with different functions that a living cell needs to cope with changes in its environment. Thus highly connected regulatory proteins usually correspond to rather complicated tasks such as e.g. the heat shock response, where about 40 chaperones are controlled by a single sigma factor, or the chemotaxis where a few regulatory proteins switch on a large number of proteins associated with flagella, flagellar motor, and sensing of the environment.

To summarize the above discussion, it is feasible that molecular networks operating in living cells have organized themselves in a particular computational architecture that makes their dynamical behavior both robust and specific. Topologically the specificity of different functional modules is enhanced by limiting interactions between hubs and suppressing the average degree 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.

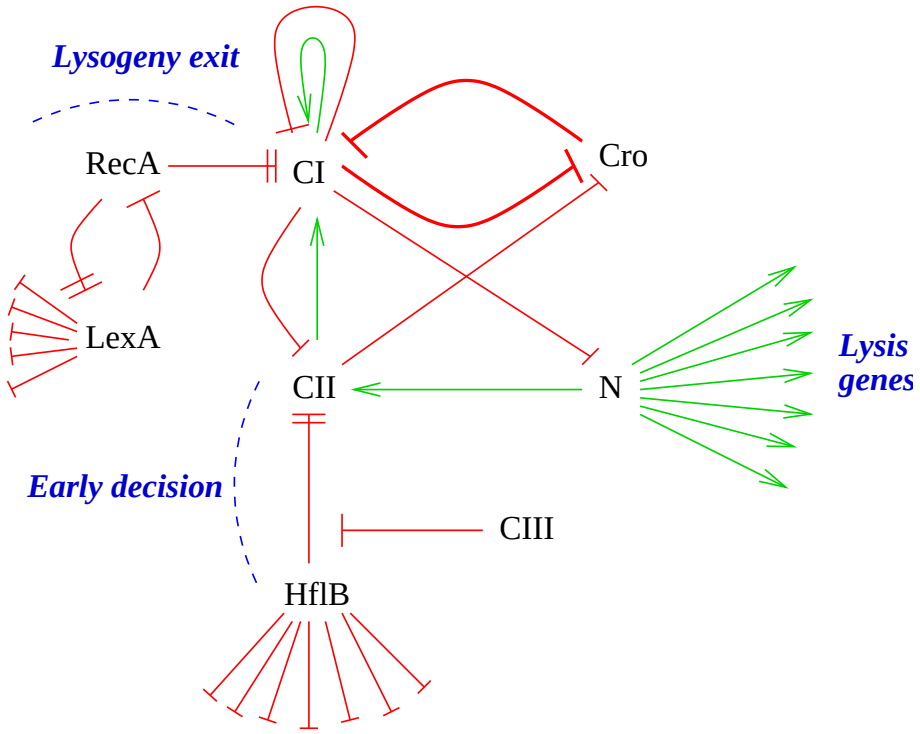


Figure 10: Lambda-phage regulatory network. The actual computation is done by centrally positioned CI, Cro and CII that have low-to-intermediate out-degree and relatively large in-degree. Their decision is transmitted to peripherally positioned, highly connected hub-proteins such as N and LexA, which in their turn broadcast it to the whole battery of response genes. As a curiosity, note that the HflB protease from E. coli's heat-shock response network interacts with the lambda-phage regulatory network. Another curiosity: the HflB directly regulates DnaK, which at least indirectly has substantial influence on the overall transcription of ribosomal RNAs of the E. coli. Thus the lambda network integrates as a small subnetwork in the overall bacterial regulatory network of E. coli. The notation used in this figure: $\uparrow$ indicates positive regulation, $\top$ indicates passive negative regulation; $\overline{\top}$ indicates active degradation through the protease activity.

# References


[1] Barabasi A.-L., Albert R. Emergence of scaling in random networks. Science 1999; 286:509-512.

[2] Jeong H., Tombor B. , Albert R. et al. The large-scale organization of methabolic networks. Nature 2000; 407: 651-654.

[3] Jeong H. , Mason S., Barabasi A.-L. et al. Centrality and lethality of protein networks. Nature 2001; 411: 41-42.

[4] As reported in the Regulon database: Huerta A.M., Salgado H., Thieffry D. et al. RegulonDB: A database on transcriptional regulation in Escherichia coli. Nucleic Acid Res. 1998; 26: 55.

[5] As reported in the YPD database: Costanzo M. C., Crawford M.E., Hirschman J.E. et al. YPD, PombePD, and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. Nucleic Acids Research 2001; 29:75-79.

[6] Stover C.K., Pham X. Q., Erwin A. L. et al. Complete genome sequence of Pseudomonas aerugiosa PA01, an opportunistic pathogen. Nature 2000; 406: 959-398.

[7] van Nimwegen E. Scaling laws in the functional content of genomes. Trends in Genetics 2003; 19(9): 479-484.

[8] Ito T., Chiba T., Ozawa R. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 2001; 98: 4569-4574.

[9] Newman M. E. J., Strogatz S. H., Watts D. J. Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E 2001; 64: 026118, pp.1-17.

[10] Maslov S. and Sneppen K. Specificity and Stability in Topology of Protein Networks. Science 2002; 296: 910-913.

[11] Maslov S., Sneppen K., Zaliznyak A. Pattern Detection in Complex Networks: Correlation Profile of the Internet. Preprint at arXiv.org e-Print archive available at http://arxiv.org/abs/cond-mat//0205379. 2002; Physica A *in press* 2003.

[12] Shen-Orr S., Milo R., Mangan S. et al. Network motifs in the transcriptional regulation of Escherichia coli. Nature Genetics 2002; 31(1):64-68.

[13] Milo R., Shen-Orr S, Itzkovitz S. et al. Network motifs: simple building blocks of complex networks. Science 2002; 298: 824-827.

[14] Early studies of these algorithms in the context of matrices were reported in: Gale D. A theorem of flows in networks, Pacific J. Math. 1957; 7:1073-1082 ; Ryser H.J. Matrices of zeros and ones in combinatorial mathematics. In: Recent Advances in Matrix Theory. Madison: Univ. of Wisconsin Press, 1964: 103-124.

For more recent references including applications to graphs see e.g.: Kannan R., Tetali P., Vempala S. Simple Markov-chain algorithms for generating bipartite graphs and tournaments. Random Structures and Algorithms 1999; 14:293-308.

[15] The set of MATLAB programs can be downloaded at http://www.cmth.bnl.gov/ maslov/matlab.htm

[16] Metropolis N. et al. J. Chem. Phys. 1953; 21:1087.

[17] Watts D. and Strogatz, S. (1998) Nature **293**, 400-403.

[18] Gavin, A.-C., Bosche M., Krause R. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 2002; 415:141-147.

[19] Uetz, P., Giot. L., Cagney G. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000; 403:623-627.

[20] Maslov S. and Sneppen K. Protein interaction networks beyond artifacts. FEBS Letters 2002; 530:255-256.

[21] Hartwell L.H., Hopfield J.J., Leibler S. et al. From molecular to modular cell biology. Nature 1999; 402(6761 Suppl):C47-C52.

[22] Albert R.,Jeong H., Barabasi A.-L. Error and attack tolerance of complex networks. Nature 2000; 406: 378-382.

[23] Vogelstein B., Lane D., Levine A.J. Surfing the p53 network. Nature 2000; 408:307-310.