Genome-wide system analysis reveals stable yet flexible network dynamics in yeast

Mika Gustafsson, Michael Hörnquist, J Bjorkegren and Jesper Tegnér

Original Publication:

Mika Gustafsson, Michael Hörnquist, J Bjorkegren and Jesper Tegnér, Genome-wide system analysis reveals stable yet flexible network dynamics in yeast, 2009, IET SYSTEMS BIOLOGY, (3), 4, 219-228.
http://dx.doi.org/10.1049/iet-syb.2008.0112

Genome-Wide System Analysis Reveals Stable yet Flexible Network Dynamics in Yeast

Mika Gustafsson[1], Michael Hörnquist[1*], Johan Björkegren [3] & Jesper Tegnér [2,3]

[1] *Department of Science and Technology, Linköping University, SE 601 74 Norrköping, Sweden.*

[2] *Division of Computational Biology, Department of Physics, Linköping University, SE 581 83 Linköping, Sweden.*

[3] *Department of Medicine, Center for Molecular Medicine, Karolinska Universitetssjukhuset, SE 171 76 Stockholm, Sweden.*

* Corresponding author, Phone: +46 11 363381, Fax: +46 11 363270, E-mail: micho@itn.liu.se

Abstract

Recently, important insights into static network topology for biological systems have been obtained, but still global dynamical network properties determining stability and system responsiveness have not been accessible for analysis. Herein, we explore a genome-wide gene-to-gene regulatory network based on expression data from the cell-cycle in *Saccharomyces cerevisae* (budding yeast). We recover static properties like hubs (genes having several out-going connections), network motifs and modules, which have previously been derived from multiple data sources such as whole-genome expression measurements, literature mining, protein-protein and transcription factor binding data. Further, our analysis uncovers some novel dynamical design principles; hubs are both repressed and repressors, and the intra-modular dynamics are either strongly activating or repressing whereas inter-modular couplings are weak. Finally, taking advantage of the inferred strength and direction of all interactions, we perform a global dynamical systems analysis of the network. Our inferred dynamics of hubs, motifs and modules produce a more stable network than what is expected given randomized versions. The main contribution of the repressed hubs is to increase system stability, while higher order dynamic effects (e.g., module dynamics) mainly increase system flexibility. Altogether, the presence of hubs, motifs and modules induce a few flexible modes, to which the network is extra sensitive to an external signal. We believe that our approach, and the inferred biological mode of strong flexibility and stability, will also apply to other cellular networks and adaptive systems.

1. Introduction

Networks have proved to be a unifying language for widely different biological systems involving, genes, proteins, metabolites and ecological food webs [1]. Cellular networks, defined by protein-protein, protein-to-gene, and metabolic interactions, determine cellular responses to input signals and govern cellular dynamics [1]. Still, though, expression data from microarrays are most common for probing into the state of cells and much analysis and network model formation are centered on this data type. This data is often analyzed by clustering over different experiments of whole-genome expression profiles, and that technique has provided important insights into gene function [2]. However, clustering alone cannot resolve gene interactions, and progress in network identification algorithms has revealed aspects of the static wiring of gene networks [3-11]. A recent study by Luscombe and colleagues [8] provided a first step towards an understanding of network dynamics by describing when different sub-networks are active during different cellular conditions in Yeast. A general review of various methods for uncovering the structure of gene regulatory networks from experimental data can be found in [12] and of graph theoretical tools for the analysis in [13,14]. Here we present an exploration of a gene-to-gene regulatory network, obtained through a network identification algorithm using gene expression data [10]. This network contains direction, strength and sign for each interaction on a genome-wide scale, which makes it possible to perform a dynamical systems analysis on the levels of genes, motifs and modules, but also on a global scale. As far as the present authors know, this is the first time such an analysis is possible and also performed for a genome-wide gene regulatory network derived from real data.

A key issue in all network model formation is the assessment of the inferred network. Since the true network seldom is known, more than to some small parts, and also this knowledge can be uncertain, it is not trivial to say whether a new edge is a false positive or a novel discovery. An experimental investigation will settle the issue with some certainty, at least for individual edges, but reliable verification on a large scale remains a challenge.[1] There is no generally accepted way to measure the quality of an inferred biological network, but at least a first step towards a commonly accepted standard was the Dialogue on Reverse-Engineering Assessment and Methods (DREAM) competition recently [17]. In the present paper, we assess on a large scale our findings by using annotations for the genes we make use of from the Gene Ontology database (GO) [18]. We also compare various statistical properties, such as degree distribution and presence of motifs, with known facts from the literature.

The rest of the paper is constructed as follows. In section 2, we recapitulate briefly the reverse engineering method and indicate how the statistical significance is ensured. Section 3 shows how the genes with high out-degrees correspond to transcription factors (TFs) and other biological meaningful entities. It also contains one of our major results, that out-hubs are often strongly repressed, as well as some statistical

[1] It might be tempting to directly compare the obtained network with others in the literature. However, before doing so, one should notice that this is non-trivial since the number, and even the interpretation, of nodes and edges often differ. Nevertheless, we compare our gene-to-gene regulatory network with some other types of regulatory networks, and it turns out that the overlaps between our network and the ones in the literature, as well as the overlaps among the ones in the literature, are small. Indeed, the overlap between our network and the one in [14] consists of seven edges, between our network and the one in [16] is one edge, and between the ones in [14] and [16] is actually zero edges.

observations on the relation between lethality and activation/repression. In section 4, we explore the existence of motifs, a study which both is in line with previous findings and uncovers some structures not presented in the literature before, to the best knowledge of the present authors. In Section 5 we study a partition of the network into modules, and find that these correspond to biological processes and are mainly self-repressing or self-activating. We also compare with direct hierarchical clustering of the expression data, and see essentially no similarity between the two partitionings, thus showing that the graph-theoretical community concept brings in a possibility for new understanding. Section 6 provides a global systems analysis, based on eigenvalues, and by adapting the definitions of stability and flexibility to the present context, we can show that the Yeast network we study is both more stable and more flexible than all networks with similar statistical properties. Eventually, the paper is concluded in Section 7 with a discussion on the relevance of the results and possible extensions of the work.

2. Network inference and statistical significance

The utilized inference algorithm is described in detail in [10] and here we only sketch the most important steps in order to make the paper self-contained. Time-course gene expression data is fitted by least squares to a set of linear ordinary differential equations of the form

$$\dot{x}_i(t) = \sum_{j=1}^{N} w_{ij} x_j(t) + \varepsilon_i(t) ,$$

where $x_j(t)$ is the gene expression at time t of gene j, N is the number of genes and $\varepsilon_i(t)$ a stochastic variable. The coefficient w_{ij} is the net effect of gene j on the transcription rate of gene i. By utilizing a LASSO-constraint [19] of the form

$$\sum_{j=1}^{N} | w_{ij} | \leq \mu_i$$

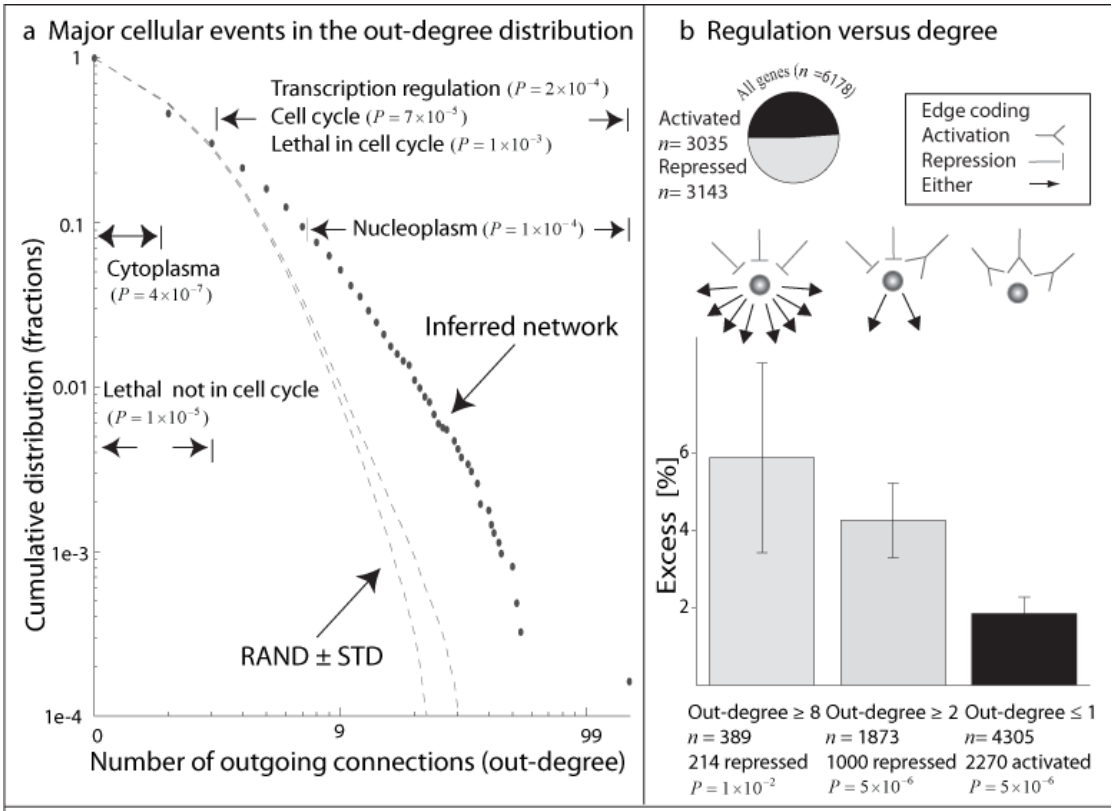
we both regularize the problem and obtain a sparse network structure. In the gene-to-gene interaction matrix w, $w_{ij} > 0$ means that gene j upregulates gene i with magnitude $| w_{ij} |$, whereas $w_{ij} < 0$ means downregulation. This algorithm is in [10] applied to the so-called extended Spellman data [20,21], consisting of 73 samples of Yeast cell-cycle data from 6178 Yeast genes (or ORFs – Open Reading Frames), resulting in the network we here explore. In [10] also all details such as missing values, estimate of time-derivatives, choice of μ_i etc., are carefully described.

All results below are evaluated against (i) shuffling the rows and columns in the array data then repeating the inference procedure (referred to as RAND) and (ii) rewiring the original network, preserving the degree distribution (referred to as REWIRED) [22]. Also some other statistical procedures are utilized occasionally, and referred to in due place.

3. Degree distribution and categorization of out-hubs

The inferred network, from [10], contains 6178 nodes (genes) and 11674 directed weighted edges (interactions), and we analyze the network statistics in detail. Figure 1a shows that the gene network has a significant (RAND) broad out-degree distribution, as previously has been observed also in protein and metabolic networks [1]. The distribution does not follow a power-law, as many previously published biological networks do (for example [23]). However, there is no theoretical justification why all networks should have this property, and there are also many examples of when this is not the case (for example [24]). We also calculate the in-

degree distribution, and obtain a quite narrow range of degrees, between one and eight, in accordance with similar calculations in [23, 24]. However, as noted in [10], this might here very well be a possible artifact of the Lasso-procedure, and we refrain from further analysis.



c Dynamics for gene groups with respect to major cellular events

Dynamic classification	n	Lethal (n= 1096)				Non-lethal (n=4597)			
		Cell cycle (n= 177)		Not in cell cycle (n=832)		Cell cycle (n=281)		Not in cell cycle (n=3030)	
		P	n	P	n	P	n	P	n
Activated	3143	0.93	81	1.00	389	0.47	144	8.3×10^{-3}	1589
Repressed	3035	9.6×10^{-2}	96	5.9×10^{-3}	443	0.58	137	0.99	1441
Activator	1434	0.27	45	0.69	188	1.1×10^{-2}	82	0.93	679
Repressor	1413	1.1×10^{-2}	54	1.00	152	0.32	68	0.22	706
Non-regulator	3331	1.00	78	6.5×10^{-4}	492	1.00	131	0.29	1645

Fig. 1:Static and dynamic network properties of the edge distribution.
(a) Cumulative distribution of out-degrees for reshuffled gene expression data (RAND) and the inferred network. Gene Ontology overrepresentation analysis for different groups of genes with P-values. (b) Mean excess of repressed genes as a function of out-degree. The bars show the excess number of repressed/activated genes (presented as fractions) from the hypergeometric distribution and the error bars corresponds to one standard deviation. (c) Table summarizing the network dynamics for gene groups divided on the basis of lethality and cell-cycle association (unknown genes are not shown). The P-values correspond to probabilities to find at least the presented number of genes with the indicated property, i.e., for example we have from the hypergeometric distribution $P = 8.3 \times 10^{-3}$ *for finding at least 1589 activated genes when we pick 3030 genes out of a set comprising 3143 activated and 3035 repressed genes. From this, one can clearly see which categories are significantly enriched and which are not.*

The out-hub categorization is obtained by rank ordering the genes according to their out-degree. We calculate the degree of overrepresentation for some biologically motivated GO-terms normally associated with high out-degree (for details see [25]). Worth noting is that the presented terms are chosen from biological knowledge and not from an exhaustive search among all terms, i.e., there is no multiple testing occurring. The analysis identifies several groups of out-hubs, e.g., genes annotated as transcription regulators, a finding consistent with previous reports that TFs can bind to several downstream genes [26]. Of special interest here are those genes associated with the cell cycle (here defined according to GO [18]), since the data come from such measurements. We observe that lethal genes are overrepresented among the cell cycle associated genes with large out-degrees (figure 1a), a finding in accordance with the previous observation that the number of connections per protein is correlated with lethality [27]. These overrepresentations are presented in figure 1a as standard *P*-values, obtained from a hypergeometric distribution, based on the annotations of the genes in GO.

To further explore the origin of lethality of Yeast genes we inspect the nature of the dynamical control exerted by the 1096 genes annotated as lethal (figure 1c). We refer to a gene with a positive sum of incoming weights as an "activated" gene, and a gene with a positive sum of outgoing weights as an "activator" gene. Corresponding definitions for "repressed" and "repressor" genes for negative sums apply. The 177 lethal genes associated with the cell-cycle are found to be repressors of downstream genes. Hence, if those repressors are knocked out, a large amount of the repression is removed from the network and an uncontrolled cascade of gene activation may occur,

causing cell death. In addition, an overrepresentation of out-going hubs may also cause an uncontrolled cascade activation of genes. To avoid such network instability it may prove beneficial for the network stability to exert strong negative regulation on precisely those genes having the largest number of out-going connections.[2] To test this hypothesis we determine the control of the out-hubs by calculating the sum of all the incoming connections. Indeed, repression is largest for out-hubs, whereas genes having few outgoing connections are not repressed (figure 1b). A similar observation about this dynamical control principle, defined by repressed and repressing hubs, has very recently been reported in Ref. 28, but is otherwise, to the best of our knowledge, unknown within systems biology. We will return to the dynamical consequences of this observation in Section 6 where we perform a system analysis.

4. Motifs

A common conjecture in present systems biology is that so-called motifs, small subgraphs consisting of few genes and of a distinct function [13,16], play an essential role in gene regulatory networks. To further analyze the network statistics we calculate all three and four gene network motifs in the network graph by applying the m-finder algorithm[3]. All results presented are statistically significant, which we here assure by only considering node-sets (i.e., motifs) found at least 20 times in the network and having large *Z*-scores (*Z*(RAND)>5 and *Z*(REWIRED)>2 [8,16]).

[2] If there are feedback loops with an even number of negative regulations in the network, i.e., effectively self-activating sub-systems, this argument is weakened. However, no such loops of reasonably short length exist in the present network.

[3] The m-finder algorithm [29] detects motifs using the adjacency matrix *a*, i.e., the matrix where the elements are $a_{ij} = 1$ if $w_{ij} \neq 0$ and zero otherwise.

First, we do not consider the signs of the interactions, and recover the previously defined motifs described in Yeast regulatory networks [8,16].[4] Feed-forward loops (FFL), bi-parallel and bi-fan motifs are overrepresented (figure 2a). In addition, our analysis reveals a previously uncharacterized 4-FFL motif.

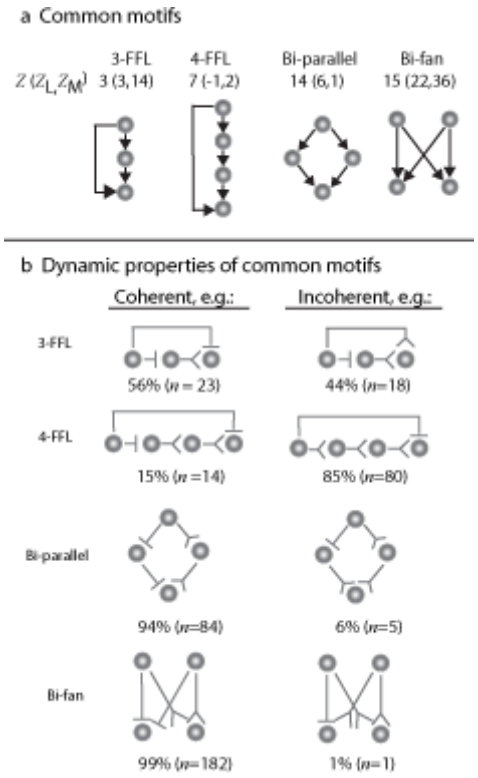


Fig. 2: Static and dynamic network motifs.
(a) Static network motifs and GO analysis. Z-scores for the inferred motifs and corresponding scores Z_L, Z_M from [8,16], respectively. The four most significant motifs with respect to both null hypotheses (REWIRED and RAND) are illustrated.
(b) Dynamic network motifs and the observed density. Same coding of the arrows as in figure 1b. The dichotomy of coherent/incoherent motifs is explained in the main text.

Second, we take into account the signs of the edges, i.e., we consider the net effect of activation and repression within a motif. Each motif can be classified as either

[4] In figure 2a, we also give the Z-values (REWIRED) for the motifs as given by [8] as Z_L and by [16] as Z_M.

coherent or incoherent. For 3-FFL, 4-FFL and bi-parallel, a motif is coherent when the two pathways have the same net effect on the target gene, and incoherent when the pathways counteract each other. For the bi-fan motif, we call a sign distribution coherent if it is possible to have states where the target genes do not receive conflicting signals, and otherwise incoherent. Note that by these definitions, the numbers of possible coherent and incoherent motifs become identical. Further, in the inferred network, the actual numbers of positive and negative edges turned out to be almost the same, which means one could expect an even distribution of coherent and incoherent motifs. Figure 2b illustrates that the Yeast network has such a distribution for the 3-FFLs motifs, but that the incoherent 4-FFL motifs are overrepresented. Incoherent FFLs have recently been shown to accelerate response-time of the gal system in E-Coli [30]. Here we find an overrepresentation of FFLs among genes annotated as being part of the cell-cycle ($P < 0.01$). This presence of incoherent motifs in the cell-cycle may therefore suggest a mixed activation and repression dynamics to reduce the response-time. The single most abundant coherent 3-FFL motif we identify is the one containing only activation (not shown) as has previously been reported by Alon and colleagues [31] derived from a literature network [16]. However, the most abundant incoherent 3-FFL motif in our hands only contains repression whereas in [31] the most abundant incoherent 3-FFL incorporated two activating and one repressing regulation. There turns out to be huge overrepresentation of coherent sign distributions among the bi-parallel and bi-fan motifs. Especially for the bi-fan, we uncover only one incoherent sign distribution. Finally, we note that the overrepresentation of coherent bi-fan motifs where the pathways are identical (which are 68% of all coherent bi-parallel motifs) may originate from gene duplication.

5. Modules

Next, we analyze network statistics beyond local motifs. Biological networks appear to be modular in nature [32,33], i.e., they are composed of more densely connected subnetworks. To determine the degree of modularity and to identify the modules, we apply a random walk Markov CLustering algorithm (MCL) [5] [34,35] to the symmetric version, $w^{(s)}$, of the weight matrix, w.[6]

The present network turns out to be highly modular (Modularity = 0.74 [37]) with 203 modules and is markedly higher than the REWIRED ensemble ($P < 10^{-80}$). For each module we submit a query to GO and obtain *P*-values for each GO-process term in the module from a hypergeometric distribution. These values are denoted as P_k^p for process term p in module k. We form a module goodness score of its biological coherence, G_k, from the logarithm of the lowest *P*-value of GO queries with at least 10% of the module members annotated[7], i.e., $G_k = -\log \min_p P_k^p$. To correct for the

[5] This is GNU-freely available software, obtained from [36], for clustering of large-scale networks, based on the steady-state flow process of biased random walkers. The efficiency of the MCL comes from its ability to produce sparse steady-state solutions from elementary matrix operations. Sparseness arises from a manipulation of the unbiased random walker algorithm, such that random walkers are biased towards already popular links. This bias introduces a free parameter, which we set to maximize the modularity [37], a goodness score indicating how well a set of modules is fitted onto a given network.

[6] This symmetrized matrix is obtained as $w_{ij}^{(s)} = | w_{ij} | + | w_{ji} |$.

[7] Similar results holds for somewhat different cut-offs. Even if we consider weighted means of the logarithmic *P*-values, such that each gene explicitly contributes by its lowest *P*-value, similar results apply.

multiple testing of querying several GO-terms, we set the null hypothesis to be the same module sizes with random members (we perform 1000 such queries for each module size). Hence we estimate an expectation value,$E(G_k)$, and a standard deviation, $\sigma(G_k)$, for the null hypothesis for each module, and transform the goodness scores into *Z*-scores as $Z_k = [G_k - E(G_k)]/\sigma(G_k)$. Each of these *Z*-scores corresponds to a *P*-value, P_k, which can be approximately found from the normal distribution. The whole network then receives a global *Z*-score as $Z = \sum_{k=1}^{n} Z_k / \sqrt{n}$, where *n* is the number of modules. This reveals the global *P*-value of the graph theoretic modules being associated to coherent biological processes to be less than 10^{-5}, thus biologically validating the inferred modular architecture. More specifically, several (17) modules contain significant groups of genes involved in the same processes ($P_k < 0.01$), e.g., biosynthesis, ribosome biogenesis and DNA replication. In figure.3 we depict the eight most significant results among the 14 modules with $P_k < 0.01$. Note, though, that one specific module normally has more than one process term associated to it.

To explore the average intra-modular communication, we assume the signs of the edges are uniformly distributed over the modules, and form the *Z*-scores

$$M_k = \frac{\sum_{i,j \in C_k} (w_{ij} - E(w) a_{ij})}{\sigma(w) \sqrt{\sum_{i,j \in C_k} a_{ij}}}.$$

Here C_k refers to the set of nodes in community *k*, a_{ij} is the adjacency matrix element (see footnote 3 above), $E(w)$ is the mean of the non-zero elements in *w*, and $\sigma(w)$ is the corresponding standard deviation. These M_k are *Z*-scores for the weighted signs

within each module. As we observe large values, we utilize a χ^2-test[8] to determine whether there is any significant tendency in the intra-modular communication, or if the internal dynamics within a module is both activating and repressing. Here, this test discards the hypothesis of a uniform sign distribution with $P < 10^{-32}$. A similar test on the inter-modular connections, i.e., the edges *between* modules, turn out to yield the result that there is no dominant sign to be found. In total there are 38 modules with a coherent dynamical action($P_k < 0.01$), for which we have almost the same number of self-activating (26) as self-repressing (17) modules.

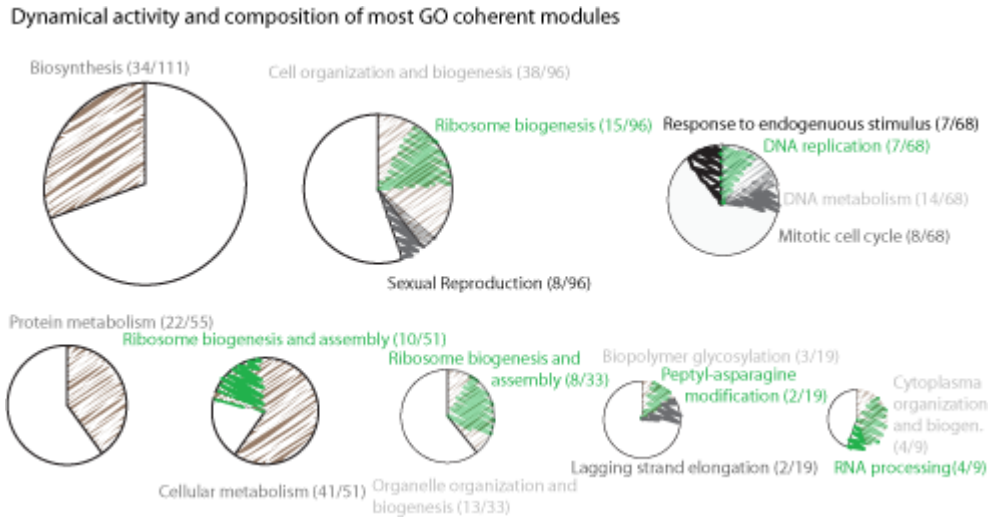


Fig. 3**: **Modular analysis of the network graph.
GO analysis of the major processes in the eight most coherent network modules. The pie charts illustrate the known module members, where the area of each chart is proportional to the number of annotated genes. The text refers to the GO-terms with least P-values and the numbers in the parentheses correspond to the actual numbers of genes in the process and in the module, respectively. Some GO-terms correspond to more than one gene, which we present as double marked areas.

[8] The legitimacy of a χ^2-test comes from the visual observation that the weights are almost normally distributed (except at zero). However, we also utilized a binomial test by simply counting the number of positive/negative interactions, with similar result.

To benchmark this partition of the network into functional modules, we also perform a hierarchical clustering of the expression data. Hierarchical clustering of whole-genome expression data has been a useful analysis technique to group genes and thereby suggest functions for uncharacterized genes. Yet, clustering does not provide or utilize any structural information about the underlying gene regulatory network, and it is important to compare the clustering with the partitioning we obtain from the inferred network. Here we choose for the clustering the same number of disjoint clusters as we obtain from the MCL-algorithm. The hierarchical clustering is in a standard form, using the correlation as distance and the furthest distance between clusters as collapsing criterion. To evaluate the similarity between the modules and clusters we utilize the similarity index I_{moved} from [38], which essentially is a normalized version of the number one gets from counting how many units have to be moved in order for the two partitionings to coincide. It turns out that the overlap between the network modules and hierarchical clusters is small, only 5%, which emphasize the novelty of the present approach. Furthermore, the same holds true for the genes contributing to the coherent processes of the modules, i.e., they are not found in similar hierarchical clusters more than is expected by random. Several modules, such as ribosome biogenesis and DNA replication, could not be detected by a regular clustering analysis since the genes with the corresponding GO terms have a low degree of correlation in their transcript activity for the present data. Clearly, the inferred network and the MCL-algorithm reveal new functional units and provide direct evidence for the relevance and existence of modules beyond the traditional clustering [39].

6. System analysis

Several authors have discussed and suggested the hypothesis that biological systems in general, and networks in particular, should have a dynamical modular organization, including motifs, leading to a stable yet flexible system [32,40,41]. Here we have shown the presence of repressed and repressing hubs, dynamical motifs, and self-activating and repressing modules. However, it is yet not clear how these properties collectively determines the overall dynamical system behaviour, and the exploration of the hypothesis "stable yet flexible" is the subject for the present section.

To study this issue in a more quantitative manner, we first need to define the entities. Although system analysis is a well established field within engineering [42], we cannot directly use the concepts from that domain, since the network we study is much more uncertain and based on data of lower quality than normal there. Nevertheless, our inferred network includes the magnitude of activation or repression for each gene-to-gene interaction, and we can explicitly calculate the eigenvalues which form the basis for any (linear) system analysis.

The degree of network stability, *S*, is here determined from the instability, *I*, which is the sum of eigenvalues, λ_i, with positive real parts. This sum corresponds to how fast a random perturbation will grow. Positive (negative) real parts of the eigenvalues correspond to unstable (stable) modes, and by summing the largest eigenvalues we can assess the degree of network instability. Explicitly, the instability is given by

$$I = \sum_{i=1}^{N_+} \lambda_i ,$$

where the eigenvalues are ordered such as $\mathrm{Re}\,\lambda_i \geq \mathrm{Re}\,\lambda_{i+1}$ and N_+ is the number of eigenvalues with positive real parts (the imaginary parts of the eigenvalues cancel each other since the secular equation here has real coefficients). System stability is then defined as

$$S = 1 - \frac{I}{I_{\max}},$$

where $I_{\max}$ is the theoretical maximum here approximated by the Gerschgorin's theorem[9].

Apart from stability, the network also has to possess flexibility, indicating the responsiveness of the system to an external signal (for a given stability). The system flexibility is here defined from the Participation Ratio, PR [44], calculated for the N_+ eigenvalues λ_i with positive real part as

$$\mathrm{PR} = \frac{\left(\sum_{i=1}^{N_+} (\mathrm{Re}\,\lambda_i)^2\right)^2}{\sum_{i=1}^{N_+} (\mathrm{Re}\,\lambda_i)^4}.$$

From this number we determine the *flexibility* as $(N_+ - \mathrm{PR})/(N_+ - 1)$, which is a normalized index between zero and unity. Explicitly, this index is large when a few eigenvalues are significantly larger than the other, which indicate the possible existence of some modes which rapidly can take the system from one state to another. Hence, for a given stability, the flexibility tells us how responsive the system is to some specific signal, internal or external.

[9] Gerschgorin's theorem states that the eigenvalues of a matrix is contained within the union of the circles with centre given by the diagonal elements and radius by the sum of absolute values of the corresponding off-diagonal element at the same row [43].

We compute the stability and flexibility for our inferred Yeast network and compare with ensembles of several randomized versions thereof. Following the arrows of figure 4, starting in the lower left corner, we have the following scenario: First, an ensemble of Erdös-Rényi (ER) like networks [1,13], having a Poisson distribution of degrees, without hubs, motifs and modules, but with the same number of nodes and directed edges with signed weights as the Yeast network, has the lowest stability and flexibility of all networks considered. Second, introducing the same degree-distribution as Yeast, but otherwise no other structure, we obtain the ensemble of REWIRED networks. Due to the wide degree-distribution, these networks contain hubs, and increase the stability and flexibility compared with the ER-networks. The stability and flexibility are further increased when modules and motifs are added to REWIRED, thus corresponding to the ensemble of Yeast Topology networks, which are the Yeast network but with randomized sign distributions of the edges. The observed network stability of the Yeast Topology network is significantly larger than both what is obtained by an array reshuffling (RAND) and by REWIRED. Worth noting is also that the ensemble of RAND networks is not markedly different from the Yeast Topology network with respect to flexibility, but still has lower stability than both REWIRED and the Yeast Topology network.

The last steps, from the Yeast Topology network to the inferred Yeast network via the Repressed Hubs, are by necessity small, we are close to the upper limit, but still of uttermost importance for the understanding of our findings of the repressing hubs and

coherent motifs and modules. The inferred distribution of activation and repression increases network stability without influencing flexibility more than slightly. A closer

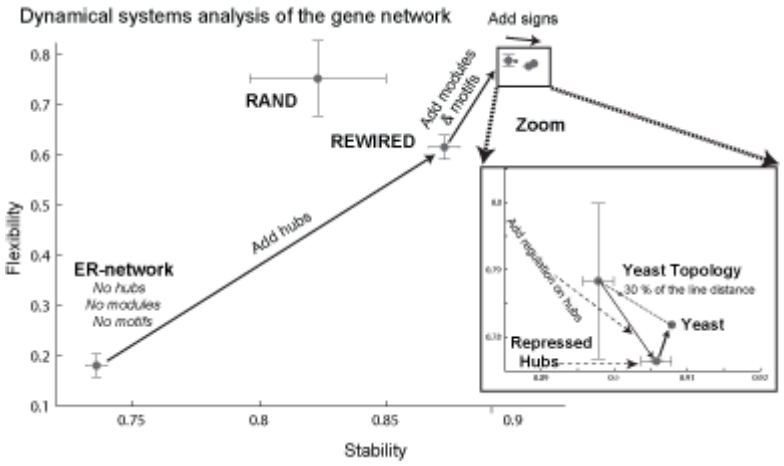


Fig. 4**: **Dynamical systems analysis of the gene network.
Stability and flexibility for the inferred network (Yeast) and several randomized versions thereof. The error bars cover two standard deviations of the ensemble networks and are obtained from repeating the design of each network (n > 300). Starting in the lower left corner of the figure, the ensemble of ER-like networks, we successively add topological and dynamical features to the network, thus obtaining new ensembles of networks more and more similar to the inferred Yeast network. It is evident that almost each of the isolated topological and dynamical features increases either stability or flexibility, or both. Aside from this exploration, we also derive the ensemble of networks obtained from totally randomized data (RAND). It is striking how this ensemble of networks is significantly less stable than both the inferred Yeast network and most of the ensembles of randomized networks. However, the RAND-ensemble turns out to be almost as flexible as the inferred Yeast network, which is unexpected but also of less relevance due to its low stability.

look, inset of figure 4, shows that the repressed hubs significantly enhance system stability, and the regulatory effect of the hubs alone comprises 75% of the increase in stability from the Yeast Topology network to the Yeast network, i.e., the point Repressed Hubs is situated only one quarter from the Yeast network along the stability axis between Yeast Topology and Yeast. As this fixing of the values of the ingoing edges to the hubs (with out-degree at least two) corresponds to 30% of all edges, we also mark in the inset of figure 4 the point representing 30% of the distance

between Yeast Topology and the Yeast network. This, together with the huge increment in stability from the ER-like networks, shows it is highly effective to concentrate on the hubs for improving stability. However, the last increase in stability comes to the expense of a decrease in flexibility. The very last step, from Repressed Hubs when all values on the edges get fixed to their values for the Yeast network, compensates this decrease somewhat and also slightly increases the stability further. Moreover, the two drastic increments in flexibility from REWIRED to Yeast Topology and also from Repressed Hubs to Yeast network in figure 4, suggest that the main reason for the occurrence of the observed complex network patterns, i.e., motifs and modules, is to produce a system responsive to selective stimuli.

This system analysis suggests that the Yeast gene network has been tuned for maximal stability while preserving the responsiveness of the network to selective external signals. That is, this arrangement may facilitate the ability of the network to rapidly switch between different dynamical states. To elucidate the function of the genes which correspond to the modes that produces large network flexibility, we eventually perform another GO analysis. We find six unique genes (YHL018W, FAA1, KCC4, HHT1, RRN5, MRPL44) in the four dominant flexible modes (i.e., the six most expressed genes in the eigenvectors corresponding to the four eigenvalues with the largest real parts) and five of those (except YHL018W, protein of unknown function) are related to primary (essential) metabolism ($P < 0.055$). This analysis therefore suggests that the regulation of these genes may be particularly important in order to control state-transitions in the network dynamics.

7. Conclusions

The next logical step in the analysis of cellular networks is the shift from describing the static topological properties to understanding the underlying dynamical principles governing network activity. Our work is one of the first attempts at a global scale in exploring dynamical network properties from signed interactions with repressing hubs, dynamical motifs and modules.

We have presented a principled statistical approach to uncover and validate the local and global structure and dynamics of cellular networks. We find that the detailed organization of activation and repression within the Yeast network is particularly important to maximize network stability and flexibility. This analysis sets the stage for understanding how biological networks are organized to balance between requirements of stability versus demands on swift responses to changes in the cellular environment. Given the statistical robustness of our derived dynamical principles we expect a similar analysis of other biological networks to reveal systems operating in a comparable dynamical regime as the Yeast network. As more quantitative high-throughput data-sets are produced we expect our approach to be widely applicable also for networks of different kinds and for other organisms. An important development in progress is how to integrate several different data-types such as gene expression measurements, transcription factor binding information, protein-protein data and sequence information into a sound statistical inference engine which thereby will increase the power of the network inference thus increasing the reliability of the reconstructed networks.

The fine-tuning of these tools will most likely be produced by work using data from model systems including Yeast and other non-mammalian cellular systems. Yet it will become increasingly important to adapt these tools for determining how the cellular dynamics is altered during human complex multifactorial diseases.

Acknowledgements

We thank Olivia Eriksson for suggesting negative repression onto hubs. We appreciate the critical comments from the computational medicine team at Linköping University and Karolinska Institutet. Financial support from the Center for Industrial IT at Linköping University (MG, MH), the Carl Trygger Foundation (MH), the Swedish research council (JB) and the foundation for strategic research (JB, JT) is hereby acknowledged.

References

1. Barabasi, A. L. and Oltvai, Z. N.: ‘Network biology: understanding the cell's functional organization’, *Nat. Rev. Genet*., 2004, **5**, pp. 101-113
2. Chua, G., Robinson, M. D., Morris, Q., and Hughes, T. R.: ‘Transcriptional networks: reverse-engineering gene regulation on a global scale’, *Curr. Opin. Microbiol*., 2004, **7**, pp. 638-646
3. Gardner, T. S., di Bernardo, D., Lorenz, D., and Collins, J. J.: ‘Inferring genetic networks and identifying compound mode of action via expression profiling’, *Science*, 2003, **301**, pp. 102-105
4. Lee, I., Date, S. V., Adai, A. T., and Marcotte, E. M.: ‘A probabilistic functional network of yeast genes’, Science, 2004, **306**, pp. 1555-1558
5. Schadt, E. E. et al.: ’An integrative genomics approach to infer causal associations between gene expression and disease’, *Naure Genetics*, 2005, **37**, pp. 710-717
6. Basso, K. et al.: ‘Reverse engineering of regulatory networks in human B cells’, *Naure Genetics*, 2005, **37**, pp. 382-390
7. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P.: ‘Causal protein-signaling networks derived from multiparameter single-cell data’, *Science*, 2005, **308**, pp. 523-529
8. Luscombe, N. M. et al.: ‘Genomic analysis of regulatory network dynamics reveals large topological changes’, *Nature*, 2004, **431**, pp. 308-312
9. Tegner, J., Yeung, M. K., Hasty, J., and Collins, J. J.: ‘Reverse engineering gene networks: integrating genetic perturbations with

dynamical modeling', *Proc. Natl. Acad. Sci. USA*, 2003, **100**, pp. 5944-5949

10. Gustafsson, M., Hörnquist, M., and Lombardi, A.: 'Constructing and Analyzing a Large-Scale Gene-to-Gene Regulatory Network-Lasso-Constrained Inference and Biological Validation', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2005, **2**, pp. 254-261

11. Thorsson, V. H., Hörnquist, M., Siegel, A.F., and Hood, L.: 'Reverse Engineering Galactose Regulation in Yeast through Model Selection', *Statistical Applications in Genetics and Molecular Biology*, 2005, **4**, (1), article 28

12. Cho, K.-H., Choo, S.-M., Jung, S.H., Kim, K.-R., Choi, H.-S., and Kim, J.: 'Reverse engineering of gene regulatory networks', *IET Syst. Biol.*, 2007, **1**, (3), pp. 149-163

13. Mason, O., and Verwoerd, M.: 'Graph theory and networks in Biology', *IET Syst. Biol.*, 2007,**1**, (2), pp. 89-119

14. Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M. , and Aravind, L.: 'Comprehensive Analysis of Combinatorial Regulation using the Transcriptional Regulatory Network of Yeast', *Journal of Molecular Biology*, 2006, **360**, pp. 213-227

15. Balázsi, G., Barabási, A.-L., and Oltvai, Z.N.: 'Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*', *Proc. Natl. Acad. Sci. USA*, 2005, **102**, pp. 7841-7846

16. Milo, R. et al.: ‘Network motifs: simple building blocks of complex networks’, *Science*, 2002, **298**, pp. 824-827

17. DREAM, Dialogue on Reverse-Engineering Assessment and Methods, 2007, project webpage: http://wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project, accessed March 2008

18. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G., ‘Gene ontology: tool for the unification of biology’, *Nature Genetics*, 2000, **25**, pp. 25–29

19. Tibshirani, R.: ‘Regression Shrinkage and selection via the Lasso.’ *J Royal Statistical Society, Series B*, 1996, **58**, pp. 267-288

20. Spellman, P. T. et al.: ‘Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization’, *Mol. Biol. Cell*., 1998, **9**, pp. 3273-3297

21. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis, R.W.: ‘A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle’, *Mol. Cell*., 1998, **2**, pp. 65-73

22. Maslov, S., and Sneppen, K.: ‘Specificity and stability in topology of protein networks’, *Science*, 2002, **296**, pp. 910-913

23. Guelzim, N., Bottani, S., Bourgine, P., and Képès, F.: 'Topological and causal structure of the yeast transcriptional regulatory network', *Nature Genetics*, 2002, **31**, pp. 60-63

24. Thieffry, D., Huerta, A.M., Pérez-Rueda, E., and Collado-Vides, J.: 'From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli', *BioEssays*, 1998, **20**, pp. 433-440

25. Eriksen, K. A., Hörnquist, M., and Sneppen, K.: 'Visualization of large-scale correlations in gene expressions', *Funct. Integr. Genomics*, 2004, **4**, pp. 241-245

26. Lee, T. I. et al.: 'Transcriptional regulatory networks in Saccharomyces cerevisiae', *Science*, 2002, **298**, pp. 799-804

27. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N.: 'Lethality and centrality in protein networks', *Nature*, 2001, **411**, pp. 41-42

28. Ma'ayan, A., Lipshtat, A., Iyengar, R., and Sontag, E.D.: 'Proximity of intracellular regulatory networks to monotone systems', *IET Syst. Biol.*, 2008, **2**, (3), pp. 103-112

29. Kashtan, N. I., Itzkovitz, S. Milo, R., and Alon, U.: 'Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs', *Bioinformatics*, 2002, **20**, (11), pp. 1746-1758

30. Mangan, S., Itzkovitz, S., Zaslaver, A., and Alon, U.: 'The Incoherent Feed-forward Loop Accelerates the Response-time of the gal System of Escherichia coli.', *J. Mol. Biol.*, 2006, **356**, pp. 1073-1081

31. Mangan, S., and Alon, U.: ‘Structure and function of the feed-forward loop network motif’, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, pp. 11980-11985

32. Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W.: ‘From molecular to modular cell biology’, *Nature*, 1999, **402**, pp. C47-52

33. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N.: ’Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data’, *Naure Genetics*, 2003, **34**, (2), pp. 166-176

34. Enright A.J., Van Dongen S., and Ouzounis C.A.: ‘An efficient algorithm for large-scale detection of protein families’, *Nucleic Acids Research*, 2002, **30** (7), pp. 1575-1584 (2002)

35. Van Dongen, S.: ‘Graph clustering via a discrete uncoupling process’, *Siam Journal on Matrix Analysis and Applications*, 2008, **30**, pp. 121-141

36. http://micans.org/mcl/ accessed February 2008, homepage of MCL by Van Dongen, S.

37. Girvan, M., and Newman, M. E.: ‘Community structure in social and biological networks’, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, pp. 7821-7826

38. Gustafsson, M., Hörnquist, M., and Lombardi, A.: ‘Comparison and validation of community structures in complex networks’, *Physica A: Statistical Mechanics and its Applications*, 2006, **367**, pp. 559-576

39. Ihmels, J. et al: ‘Revealing modular organization in the yeast transcriptional network’, *Nature Genetics*, 2002, **31**, pp. 370-377

40. Kitano, H.: ‘Computational systems biology’, *Nature*, 2002, **420**, pp. 206-210

41. Csete, M. E., and Doyle, J. C.: ‘Reverse engineering of biological complexity’, *Science*, 2002, **295**, pp. 1664-1669

42. Ljung, L., and Glad, T: ‘Modeling of Dynamic Systems’ (Prentice Hall, Upper Sadle River, N.J., 1994)

43. Råde, L., and Westergren, B.: ‘Beta, Mathematics Handbook’ (Studentlitteratur, 1988, 2nd edn. 1990)

44. Wegner, F.: ‘Inverse Participation Ratio in 2+eps Dimensions’, *F. Phys. B*, 1980, **36**, pp. 209-214