Analysis of Biological Networks: Network Integration *

Lecturer: Roded Sharan
Scribe: Shelly Mahleb and Benny Davidovich

Lecture 13, January 18, 2007

Introduction

The integration of different types of networks over the same set of elements enables the identification of motifs/modules supported by several data types, studying inter-relations between different networks and predicting protein function and interaction. Network motifs are characteristic network patterns consisting of both transcription-regulation and protein-protein interactions that recur significantly more often than in random networks. Large-scale studies have revealed networks of various biological interaction types, such as protein-protein interaction, genetic interaction and transcriptional regulation. A system-level understanding can be achieved by providing models of both the molecular assemblies involved and of the functional connections between them.

1 Network Motifs

Network motifs are patterns of interconnections occurring in real networks at numbers that are significantly higher than those in randomized networks [2]. Network motifs help to uncover the structural design principles of networks, and in studying the relations and functions of genes and the proteins they encode.

Various processes within the cell are mediated by protein-protein interactions (PPIs) and transcriptional interactions (TIs). Those interactions are modeled as interaction networks, and there are more than a dozen methods to detect them. As a result, at present there is a number of different interaction networks available for each sequenced organism. However, the networks generated by different methods are often not superposable in any obvious way. Moreover, analyzing transcriptional networks and PPIs networks separately hides the full complexity of cellular processes, because many of them involve combination of these two types of interactions. Integrating these different networks in order to arrive at a statistical summary of which proteins work together within a single organism can help detect linkages that would have been missed if only one kind of interaction was studied. It can also help strengthen the confidence of known linkages.

Searching for network motifs in an integrated network requires expanding the basic definition of network motifs to patterns that consist of two (or more) interaction types.

1.1 Transcriptional-PPI Networks

The work of Yeger *et al.*[10] analyzes motifs in an integrated transcriptional and protein-protein interactions network in yeast. The resulting network is described as a graph; each node represents both a protein and the

*Based on last's year scribe, written by Andrey Stolyarenko, Yaron Ardenboim and Hadas Zur

gene that encodes it. A PPI is represented by a bidirected edge connecting the interacting proteins and is colored in black; a TI represented by a directed edge pointing from the transcription factor to its target gene and is colored in red (see Figure 1).

Figure 1: An example of an integrated PPI-TI network.

A motif is identified as pattern that occurs at least five times, and statistically significantly more than in randomized networks. In order to generate randomized networks the approach of Shen-Orr *et al.* [6] involving one type of connection was extended. To this end the following terms were defined:

1. *Extended degree of a node* - the number of edges per edge type that point to or from a node. Two nodes have the same extended degree if they have the same number of ingoing and outgoing edges for each edge type. This reflects the local connectivity of a node.

2. *Edge profile of two nodes* - the set of edges connecting the two nodes, each edge is described by its type and direction. This provides a local measure of the relation between two nodes.

The randomization is done in a way that the extended in-degree and out-degree, and the edge profile of every two nodes are preserved, as demonstrated in Figure 2.

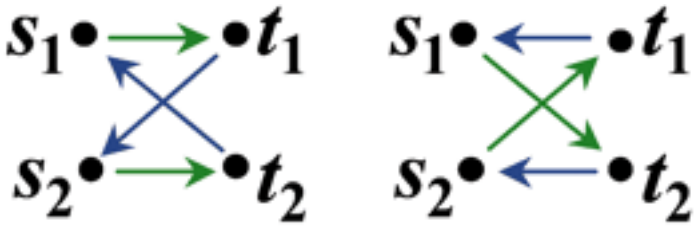


Figure 2: Network randomization. Right: Before randomization. Left: After randomization. If edge profile $(s1, t1)$ = edge profile $(s2, t2)$ and edge profile (s1,t2) = edge profile $(s2, t1)$, then edges can be switched as exemplified. For clarity, each edge color represents a type of edge profile. Note that if $(s1, t1; s2, t2)$ are switchable, then so are $(s1, t2; s2, t1)$, $(t1, s1; t2, s2)$, and $(t2, s1; t1, s2)$. Switchability is considered only for cases in which all four nodes are distinct and at least one edge profile is not empty.

Using the randomization technique described above, the resulting random network will preserve all two-nodes motifs. Therefore, for detecting such motifs, randomization is applied separately to each network, and then the two networks are merged. To overcome the noise of experimental interaction data collected via high-throughput methods, a stringent data set containing 3,183 interactions between 1,863 proteins was generated: PPIs were included if detected by at least two different experimental studies (different yeast two-hybrid methods were considered as different studies). TIs were included if detected by methods other than genome-wide experiments. The resulting network was denoted as the *stringent* network. The robustness

of the results was confirmed by performing the same analysis on a network containing all 12,413 experimentally identified interactions between 4,651 proteins (the non-stringent network). There are five possible two-protein connected patterns (Fig. 3). The statistical significance of the five patterns was assessed by comparing the number of their occurrences in the interaction network to that expected at random. Only one of the patterns, the mixed-feedback loop comprising one PPI edge and one TI edge was found significant. In this motif, protein P regulates gene g at the transcription level, and the product of gene g interacts with P at the protein level.

Figure 3: All possible interaction patterns between two connected proteins. D - feedback loop is the only two-nodes motif found.

There are 13 possible three-protein connected patterns with a single type of directed interaction, where only five of them were represented on the network. For two types of interactions, such as TI and PPI, the number of possible patterns rises to 100. Of the 100 possible three-protein connected patterns, 29 different patterns occurred in the stringent network. Only five of these occurred significantly more often than expected in random networks (P-value < 0.001) and thus constitute three-nodes motifs. These five patterns were also found to be significantly over-represented in the non-stringent network (Figure 4).

		No. of occurrences					
		Stringent network			Nonstringent network		
Motif*	Illustration†	*N* real	*N* rand ± SD	*z* score	*N* real	*N* rand ± SD	*Z* score
A. Protein clique		1,293	14 ± 3.8	332.7	2,016	87 ± 10.8	177.9
B. Interacting transcription factors that coregulate a third gene		243	2.4 ± 2.1	115.9	476	9.6 ± 7.8	59.7
C. Feed-forward loop		83	26 ± 6	9.5	994	473 ± 36.7	14.2
D. Coregulated interacting proteins		66	2 ± 1.4	46.5	285	107 ± 10.1	17.7
E. Mixed-feedback loop between transcription factors that coregulate a gene		46	2.7 ± 1.6	26.3	118	8.2 ± 5.4	20.3

Figure 4: Significant three-nodes motifs

Conclusions

At the level of two-protein patterns, the mixed-feedback loop with one PPI edge and one TI edge (as shown in Fig 3D)was found to be highly significant motif as an oppose to feedback loop with two TI edges (Fig 3C) that was as common as in randomized network. This might be a result of the response time - each

transcriptional edge causes a delay of approximately one life time of the protein product. At the three-protein level the clique configuration that represent three PPIs was the most common; it represents complexes of interacting proteins that work together as a multi-component machine. When looking for a four-protein motifs, almost all of them contained one or more of the three-protein motif with a dangling forth node or as a combination of two or more three-protein motifs. This would have been resolved by factoring out motifs of smaller sizes, however, the randomization technique only factors out two-nodes motifs, hence the results may be too optimistic. Multiple testing issue was not treated at all, as the fact that there are 100 possible three-gene connected patterns should be considered when setting the *P*-value threshold.

1.2 Motifs and Themes

The work of Zhang *et al.*[11] integrated 5 types of yeast networks in order to search network motifs and themes. Nodes in the network represent genes or proteins, and differently colored links represent different biological interaction types. The following networks have been integrated:

1. *P* - Stable protein interactions defined by shared membership in a protein complex.
2. *S* - Synthetic sick or lethal (SSL) interactions derived from synthetic genetic array (SGA) analysis.
3. *H* - Protein sequence homology relationships from a genome-wide BLAST search.
4. *R* - Transcriptional regulatory interactions from a genome-wide chromatin immuno-precipitation (ChIP) study.
5. *X* - Correlated mRNA expression relationships derived from microarray data.

This collection of data resulted in a single integrated network involving 5,831 nodes and 154,759 links.

Network themes are classes of higher-order recurring interconnection patterns that encompass multiple occurrences of network motifs [11]. Network themes can be tied to specific biological phenomena and may represent more fundamental network design principles. Examples of network themes include a pair of protein complexes with many inter-complex genetic interactions which represent the 'compensatory complexes' theme. Thematic maps are networks rendered in terms of such themes which can simplify an otherwise confusing tangle of biological relationships.

Zhang et *al.* searched for 3-motifs defined by a single type of link between each pair of nodes. After the 3-motifs were found the theme that was generated in the integrated network by each motif was observed (Figure 5). Most motifs can be explained in terms of higher-order structures, or network themes, which are representative of the underlying biological phenomena. These motifs were classified into seven sets, as shown in Figure 5a-g.

The first motif set contains the transcriptional feed-forward motif (Figure 5a). Because transcriptional regulation links often overlap co-expression links, another motif composed of two genes with correlated expression that are also indirectly connected by transcriptional regulatory links through an intermediate gene was added to this set. Most gene triads matching the feed-forward motif belong to such clusters, implying a 'feed-forward' theme - a pair of transcription factors, one regulating the other, and both regulating a common set of target genes that are often involved in the same biological process.

The next set contains 'co-pointing' motifs, in which a target gene is regulated by two transcription factors that interact physically or share sequence homology. These co-pointing motifs reflect the fact that two transcription factors regulating the same target gene are often derived from the same ancestral gene, or function as a protein complex. The authors found that these motifs also overlap extensively, forming a

co-pointing theme, in which multiple transcription factors, connected to one another by physical interaction or sequence homology, regulate a common set of target genes (Figure 5b).

A third set of motifs contains two targets of the same transcription factor bridged by a link of correlated expression, protein-protein interaction, or sequence homology. These motifs indicate that transcriptional co-regulation is often accompanied by co-expression, membership in the same protein complex, or descent from a common ancestor, and suggest a 'regulonic complex' theme in which co-regulated proteins are often components of a complex or related by gene duplication and divergence (Figure 5c).

The fourth motif set consists of four three-node motifs each containing protein-protein interactions or correlated expression links. Protein-protein interaction is known to correlate positively with co-expression, and proteins corresponding to these motifs often reside in the same complex. Thus, motifs within this set are likely to be signatures of a 'protein complex' theme (Figure 5d).

The fifth motif set contains three-node motifs linked by SSL interaction or by sequence homology. In the SSL network, neighbors of the same gene often interact with one another. This translates into a triangle motif of three SSL links. Furthermore, homology relationships are often transitive (that is, if gene A is homologous to gene B, and gene B is homologous to gene C, then gene A is often homologous to gene C). These phenomena, combined with the fact that genes sharing sequence homology have an increased tendency to show SSL interaction, suggest an underlying theme of the neighborhood clustering in the integrated SSL/homology network: SSL or homology neighbors of one node tend to be linked to one another by SSL interaction or sequence homology (Figure 5e).

The sixth motif set describes network motifs containing two nodes linked either by SSL interaction or by sequence homology, with a third node connected to each of them through protein-protein interaction or through correlated expression. This can be generalized to a network theme of a protein complex with partially redundant or compensatory members (Figure 5f).

The seventh motif set that was found was particularly interesting. Motifs in this set contain two nodes linked by protein-protein interaction or correlated expression, with a third node connected to both either by SSL interaction or by sequence homology. Considering previously observed correlations between protein-protein interaction and co-expression and between SSL interaction and sequence homology, these motifs indicate that members of a given protein complex or biological process often have common synthetic genetic interaction partners (Figure 5g).

Thematic Maps

Themes were generated by collapsing protein complexes into nodes (see Figure 6). Between-pathway and co-regulated complex themes were tested and found to be prevalent.

In order to identify additional pairs of protein complexes with overlapping or compensatory function, a map of the network in terms of the 'compensatory complexes' theme was made. This map can also serve as a guide to 'redundant systems' within the integrated network, wherein two complexes provide the organism with robustness with respect to random mutation when each complex acts as a 'failsafe mechanism' for the other. To generate a thematic map of compensatory complexes, pairs of protein complexes with many inter-complex SSL interactions were searched. For each pair of protein complexes, the number of links between them was assessed and the significance of enrichment was calculated . Among the 72 complexes examined, 21 pairs of complexes showed significant enrichment ($p \leq 0.05$) for inter-complex SSL interactions.

All possible pairings of a transcription factor with a particular protein complex (together, a 'TF-complex pair') were examined. The integrated network of stable protein-protein interactions and transcriptional interactions was reduced to one in which nodes are either transcription factors or complexes and links indicate transcriptional regulation (with multiple links allowed between a pair of nodes). For each TF-complex pair, the number of links between them was calculated, and the significance (the probability of obtaining at least the observed number of links if each transcription factor would choose its regulatory targets randomly) was

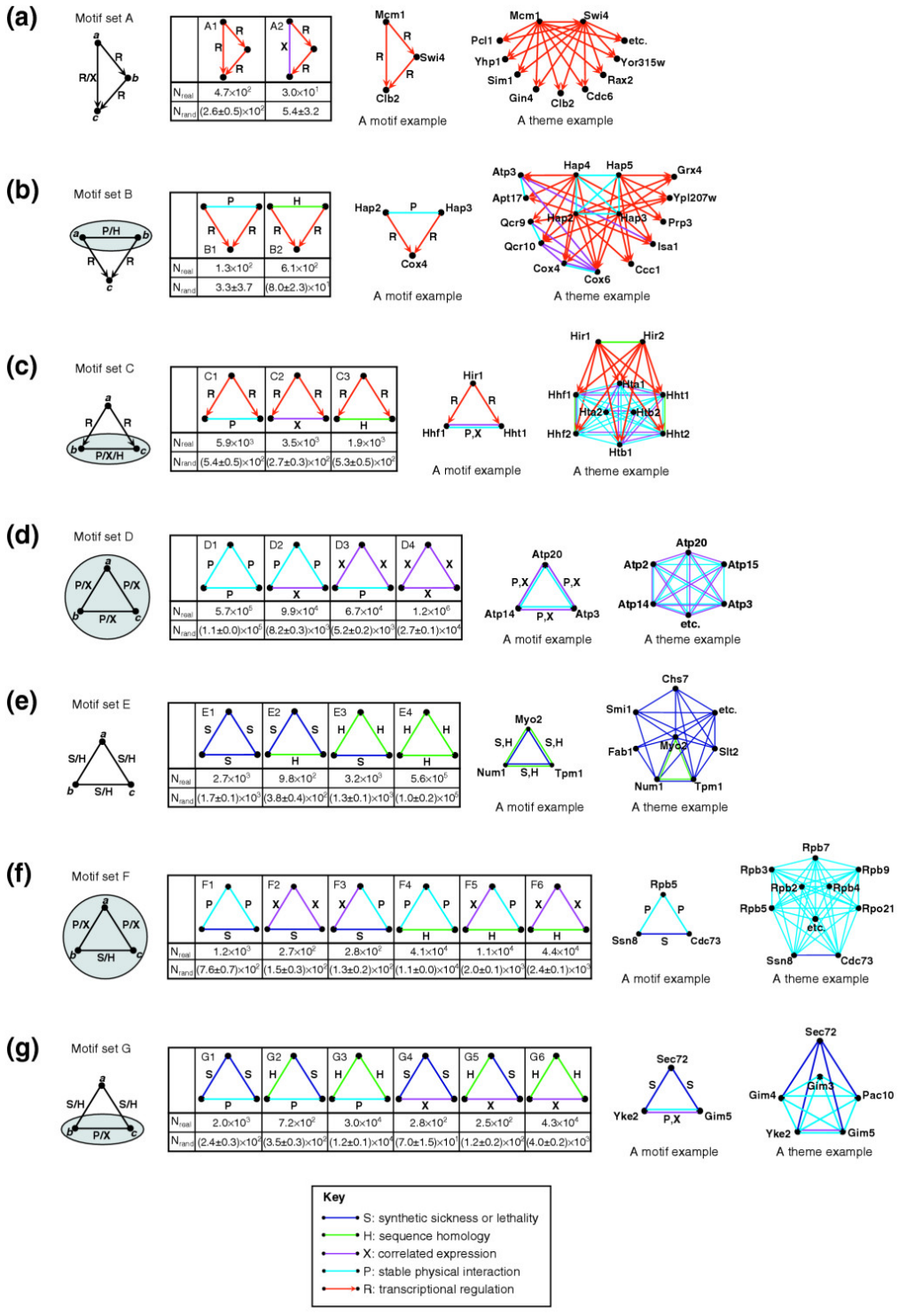


Figure 5: three-Motifs and examples of the themes they generate. (a) Feed forward (b) Co-regulating (c) Co-regulated complex (d) Protein Complex (e) Clustering (f) Complex with redundant or compensatory members (g) Between-pathway GIs .

assessed. A total of 91 TF-complex pairs showed significant enrichment ($p \leq 0.05$) for transcriptional regulation links. These significant TF-complexes relationships can also be viewed as a network whose nodes are transcription factors or complexes and whose links represent TF-complex pairs with significantly enriched transcription regulation (see Figure 6a). Many of the links connect transcription factors and protein complexes involved in the same biological process, and complexes of related function are often connected to the same transcription factor (Figure 6b).

2 Module Finding

A *module* is a set of proteins performing a distinct biological function in an autonomous manner. The genes in a module are characterized by coherent biological behavior with respect to the data at hand. Data integration allows overcoming noise and incomplete information problems and provides more complete information on the module activity and regulation. There are two common techniques for data integration: (1) Identical modeling of all data types, e.g., searching for a set of proteins that are dense with respect to interactions within the complex itself. (2) Different models for different data types. The latter is illustrated by the work of Kelley and Ideker [4]. They demonstrate that by combining genetic interaction data with information on physical interactions, it is possible to uncover physical mechanisms behind many of the observed genetic effects.

In the work of Tan *et al.* [7] PPI and TI data were integrated in a single model which simultaneously detects protein complexes and their transcriptional regulators. Their approach is based on integrating the protein protein and transcriptional interaction networks of a species, and searching for sets of proteins that densely interact in the PPI network and whose gene promoters are targeted by the same transcriptional factors in the TI network. Such protein sets are termed as coregulated protein clusters. A log likelihood ratio score was defined and a search for protein clusters was done using a greedy approach that starts from high scoring seeds and refine them by using local search. The resulting score would be compared to that of random clusters.

At first, 72 significant co-regulated protein clusters were identified, as shown in Figure 7 and in Figure 8. This results a bipartite graph, that has transcription factor on one side and protein cluster members on the other. Comparison of co-regulated clusters that were tested for functional enrichment, expression coherency and conservation coherency of their members.

The result was compared to that of a collection of 452 protein clusters inferred by using the PPIs data only, ignoring the TI data. They also included in the comparison two collections of complexes derived by co-IP experiments. They found that the coregulated clusters exhibited substantially higher expression coherency and conservation coherency levels than the experimentally derived complexes and the baseline clusters. Furthermore, 100% of the clusters were functionally enriched, slightly higher than the baseline clusters (99%) and markedly higher than the experimentally derived complexes .

Comparing these results with previous work, it is clear that an integrated search of the PPIs and TIs networks finds significant signal in the data, whereas a fixed search of TI interactions versus PPIs complexes yields a much weaker association. Only 9 of the 78 PPIs complexes were found to have a significant association to a transcription factor. Of the 72 coregulated clusters that were identified, 50 (69%) had no overlap with any of these 9 PPIs complexes, emphasizing the usefulness of this approach in identifying previously uncharacterized regulated complexes.(see Figure 9).

2.1 Interactions Prediction

A coregulated protein cluster, which involves direct transcriptional regulation of some cluster members by a specific TF (or more than one), supports the prediction that the same TF directly regulates other members

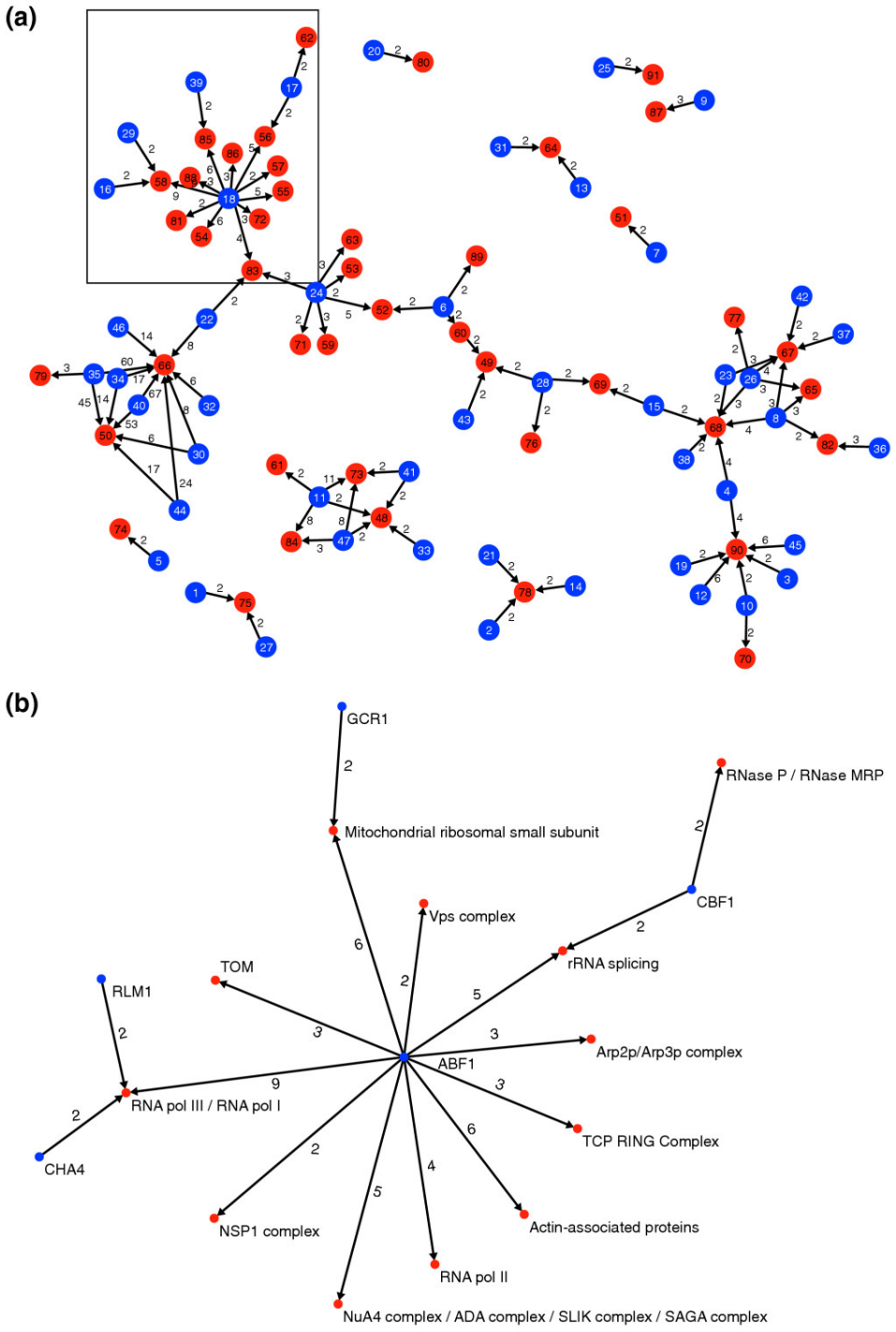


Figure 6: A thematic map of regulonic complexes. (a) Blue nodes represent transcription factors, red nodes represent protein complexes, and a link is drawn between a transcription factor and a protein complex if the promoters of a significantly large number of complex members are bound by the transcription factor. (b) An enlarged region of the regulonic complex map in (a). Links between transcription factors and the complexes they regulate are labeled with the numbers of supporting interactions in the transcription regulation network.

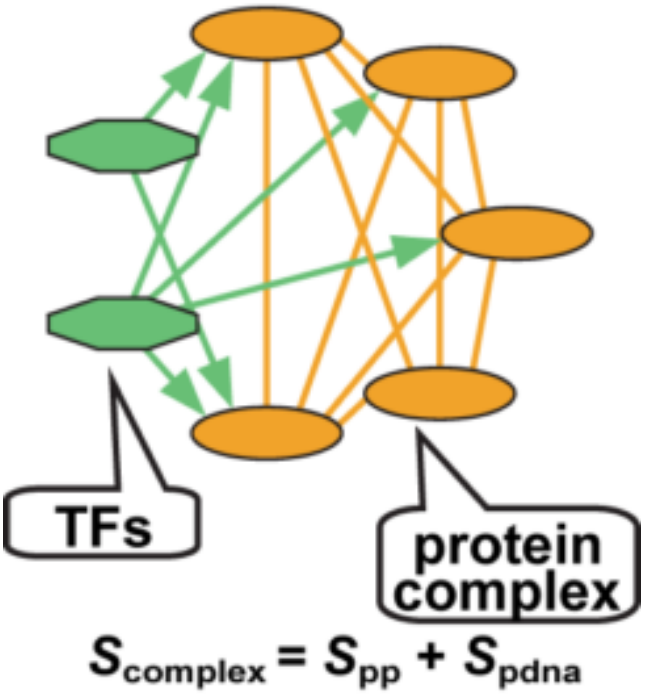


Figure 7: A typical coregulated cluster and its scoring scheme.Orangeovals, protein cluster members;blue octagons, TFs;orange lines, PPI; blue arrows, TI.

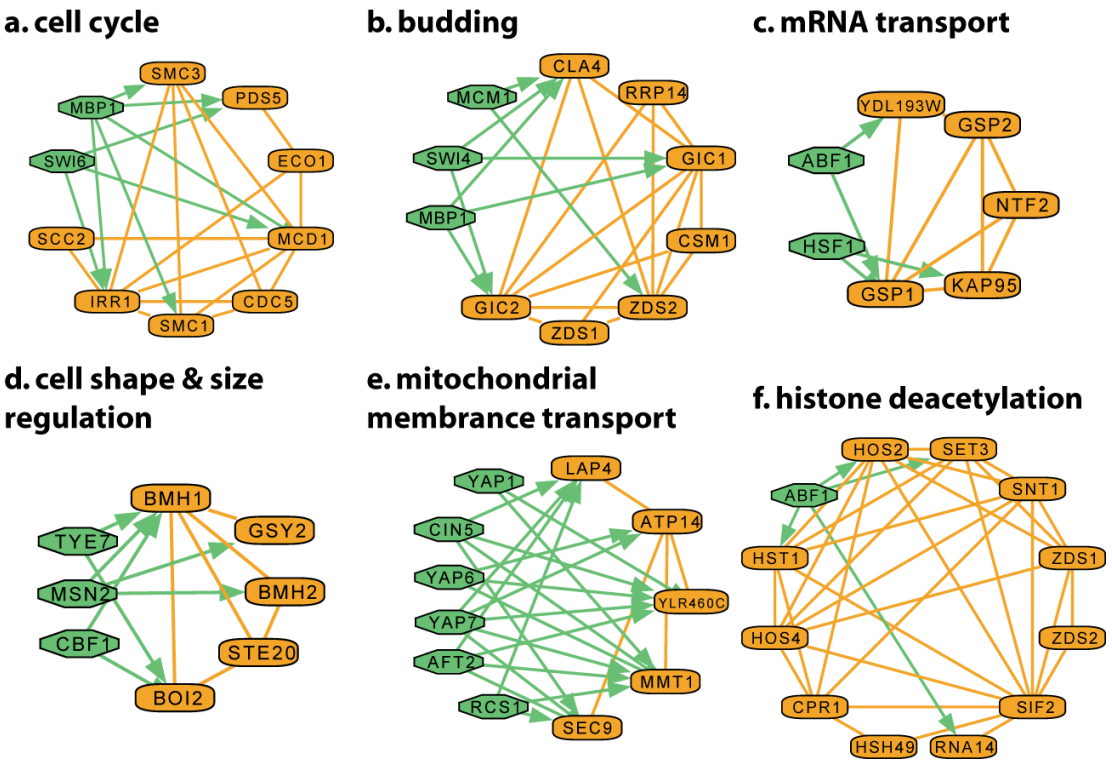


Figure 8: Representative examples of coregulated protein clusters. Shown are enriched Gene Ontology (GO) biological processes ($P < 0.05$) of clusters: cell cycle (a); budding (b); cytoplasmic transport (c); cell shape and size regulation (d); mitochondrial membrane transport (e); histone deacetylation (f).

	Complex source	GO enrichment	Expression coherency	Conservation coherency
MS-derived complexes	Ho et al. (21)	61%	8%	24%
	Gavin et al. (22)	77%	9%	36%
Protein clusters	Current study	99%	26%	22%
Co-regulated clusters	Current study	**100%**	**45%**	**59%**

Figure 9: Validation of yeast clusters by functional enrichment, expression coherency, and conservation coherency of their members

of the cluster. To prioritize these predictions, the extent to which the predicted TF targets had correlated expression and phylogenetic conservation with the respective TF was assessed, as well as the presence of known TF-binding sites in their promoters. All the measures were combined within a logistic regression classifier to assign a quantitative confidence score to each potential transcriptional interaction. This classifier attained high sensitivity (82%) and specificity (91%) levels in 10-fold cross validation (Figure 10). Overall, combining the classifier scores with the coregulated cluster information, 120 previously uncharacterized transcriptional interactions involving 23 TFs and 99 protein cluster members were predicted. To evaluate the accuracy of these predictions, 12 predicted transcriptional interactions for the TF Rpn4 were tested experimentally. Previous studies have established that Rpn4 could be activated by multiple types of cellular stresses, including heat shock. The expression profiles of wild type and rpn4 gene deletion strains under heat-shock-induced stress were compared. Seven (RPN5, RPN12, CCT8, PRE5, RPT5, PRE10, PRE2) of the 12 newly predicted Rpn4 targets exhibited differential expression ($P < 0.05$). This fraction (58%) was significant when compared with the fraction of differentially expressed genes overall (11%); it was also much higher than the one attained when predictions were made by the classifier alone (19%) (see Figure 11 and Figure 12).

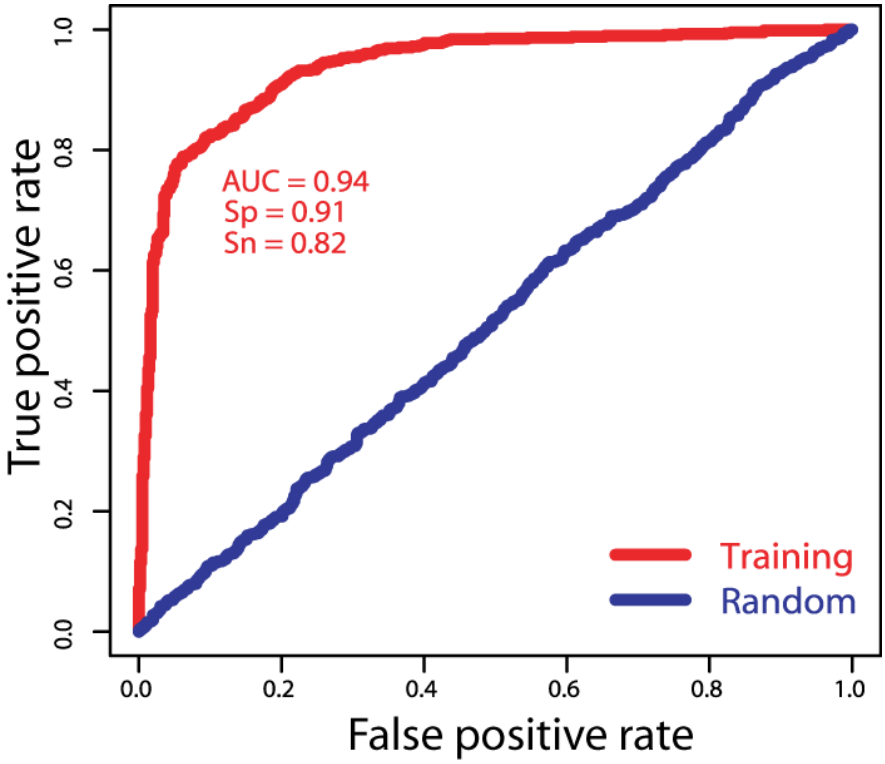


Figure 10: Transcriptional interaction prediction in yeast. Receiver operating characteristics curve of the logistic regression classifier. AUC, area under the curve; Sn, sensitivity; Sp, specificity.

2.2 Caenorhabditis Elagans Development

This section describes a work that identifies the cellular machines involved in the development process of worm [3]. It focuses on 661 genes implicated in early embryogenesis (EE genes) and using phenotype data: 45-long vectors of defects caused by silencing (RNAi). Integration is done over protein-protein, co-expression, and phenotype similarity interactions.

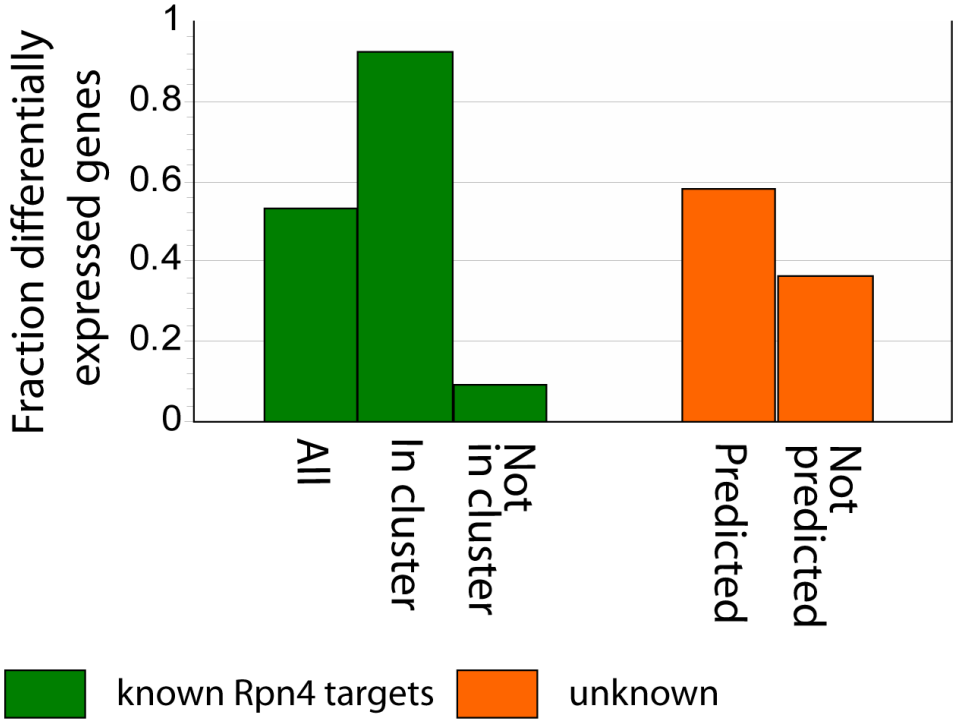


Figure 11: Fraction of differentially expressed genes in various gene sets. Green, genes bound by Rpn4 . Orange - genes in cluster models but not bound by Rpn4.

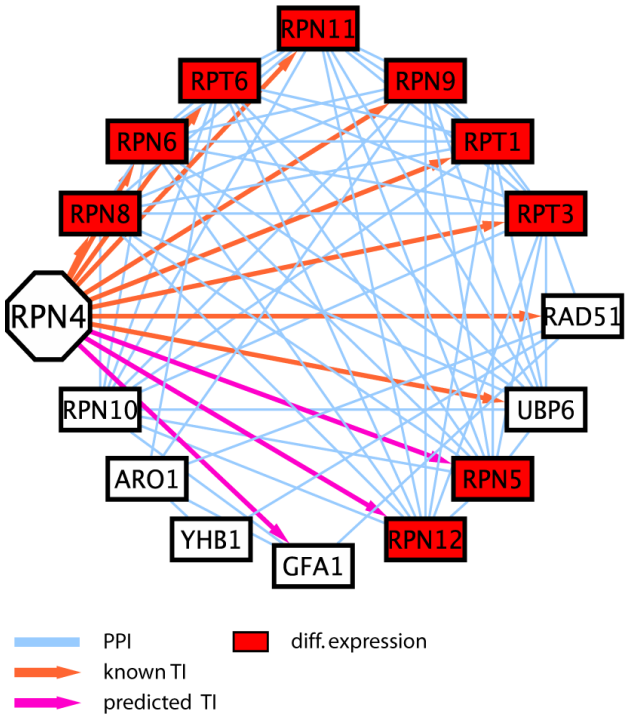


Figure 12: An example of a predicted cluster regulated by Rpn4. Orange arrows, known Rpn4 TIs. Purple - newly predicted Rpn4 TIs. Shades of red represent P values (≤ 0.05) for differential gene expression.

Phenotype Clustering

To group genes by phenotypic similarity hierarchical clustering was preformed, and the genes were organized into 23 phenotype clusters (PC) (Figure 13). It was shown that clusters tend to be significantly enriched for specific gene functions. These results suggest that phenotypes derived from RNAi data represents a reasonable way to compare genes quantitatively.

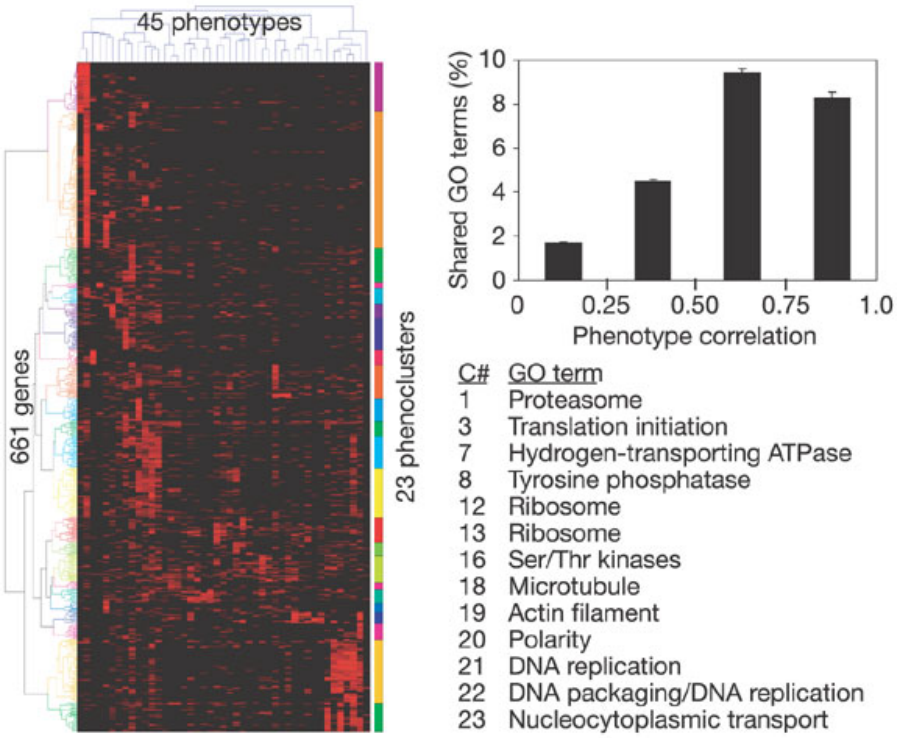


Figure 13: Phenotype hierarchial clustering. Observation shows that phenotype correlation implies GO enrichment in a cluster. On the left - the resulting gene expression of 661 genes under different conditions. The same intensity of red mark implies on similar expression level.

Data Correlations

EE genes were found to be significantly PPI connected (Figure 14). There are 513 interactions compare with about 120 in random. Genes within the same phenotype cluster were found to be significantly PPI connected as well. Phenotypic and expression similarity are correlated, particularly for interacting pairs, both increase with network proximity.

Global correlations between transcriptome profiling and interactome data sets have been used to derive network graphs that combine similarity relationships from transcription profiling with physical interactions between proteins. Suggestive correlations between interactome or transcriptome data and phenotypic data sets support the notion that these three types of data might complement one another in predicting functional relationships.

- Figure 15 shows that genes within the same phenotype cluster are significantly PPI connected.
- Figure 16 shows that proportion of PPI rises with phenotype and expression correlation.

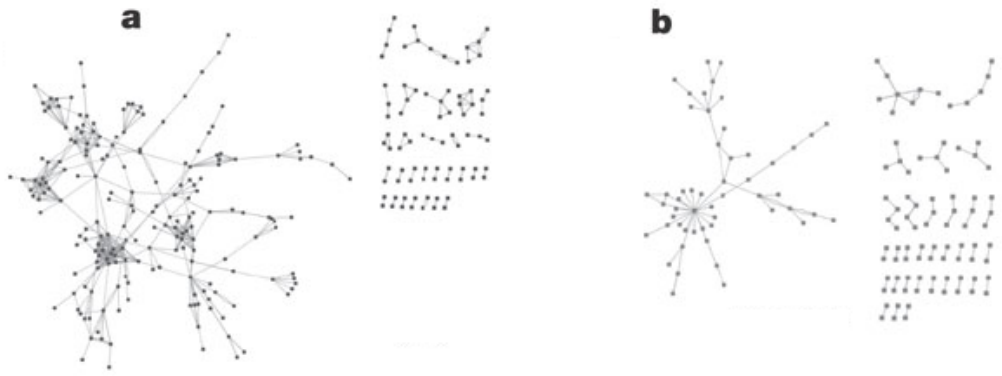


Figure 14: EE network VS. Random network. (a) EE Network 513 edges cc=0.53 (b) Random Network 120±20 edges cc=0.05.

- Figure 17 show that phenotypic and expression similarity are correlated, particularly for interacting pairs. The average correlation is decreasing significantly after a network distance of 2.

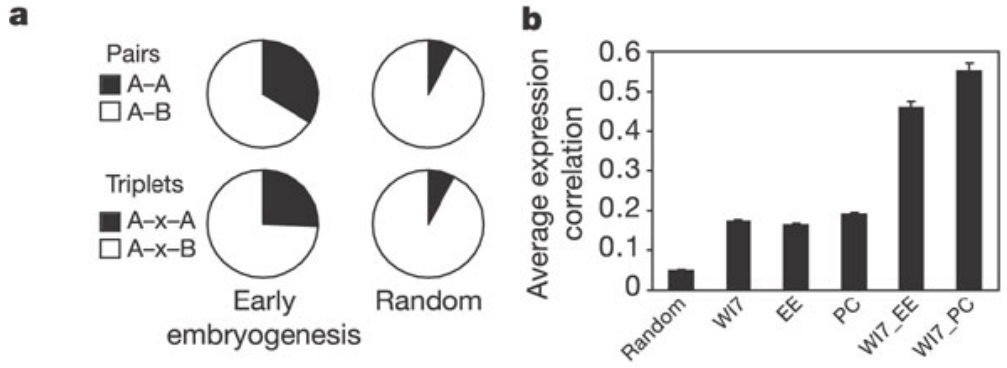


Figure 15: Data Correlation. (a) The early embryogenesis interactome subnetwork is enriched for interactions within the same phenocluster (A-A and A-x-A) relative to interactions between phenoclusters (A-B and A-x-B) (b) Interacting proteins (WI7), random early embryogenesis (EE) pairs, intra-phenocluster early embryogenesis pairs (PC), pairs of interacting early embryogenesis proteins (WI7_EE) and interacting early embryogenesis proteins from common phenoclusters (WI7_PC) all show higher expression correlation than random pairs.

Integrated Network

The integrated early embryogenesis network-joining all 661 early embryogenesis proteins by the union of all three types of relationship (See Figure 18) contains a main component with 31,173 edges characterized by an average of 0.9, 5.0 and 44 edges per node for protein interaction (Int), expression similarity (Tr) and phenotypic similarity (Ph), respectively. In this network, the number of protein pairs with doubly supported edges is significantly higher than expected by chance ($P = 10^{-34}, 10^{-61}$ and 10^{-109} for Int-Tr, Ph-Tr and Int-Ph associations, respectively).

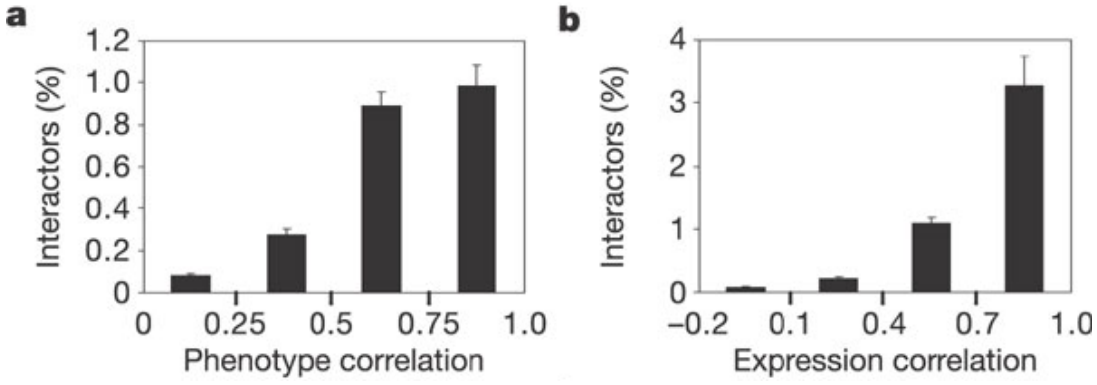


Figure 16: The proportion of physical interactions increases with phenotypic(a) and expression correlation(b)

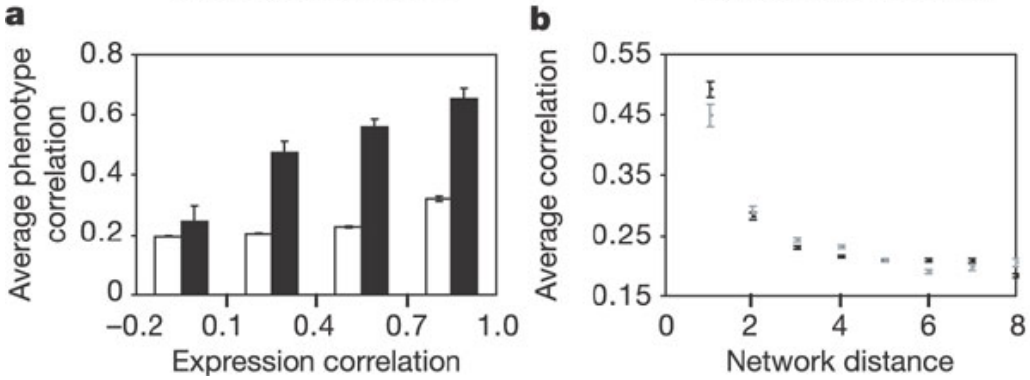


Figure 17: (a) Early embryogenesis genes with similar expression profiles are more likely to share similar RNAi phenotypes. All early embryogenesis gene pairs (open bars) and interacting early embryogenesis proteins (filled bars) were binned by expression correlation and plotted against average phenotypic correlation. (b) Phenotype and expression correlation increase with interactome proximity. Average phenotype (black) and expression (grey) correlation decrease for early embryogenesis protein pairs as their distance (shortest path) increases.

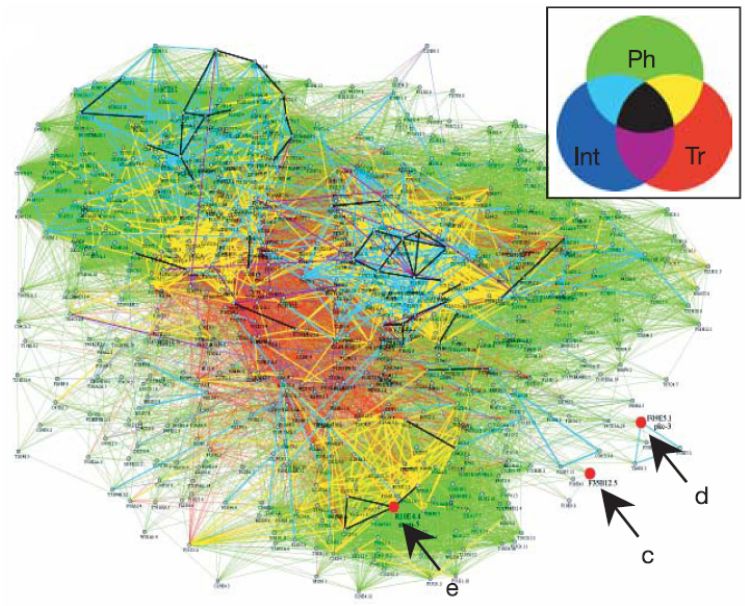


Figure 18: Entire early embryogenesis network graph. A Venn diagram (inset) shows the color system for labeling edges based on available evidence: phenotypic profiling similarity (Ph; green), expression profiling similarity (Tr; red), physical interaction (Int; blue) and overlapping combinations of data types (intersecting regions)

Function Sharing

Functional analysis of early embryonic networks was used to generate predictive models. To assess the predictive value of the early embryogenesis network on a global scale, they analyzed the individual and combined networks for their ability to predict a specific shared function between two linked gene pairs using GO annotations. The table below is an analysis of shared GO functional annotations within EE networks, considering only gene pairs for which both members have some GO annotation, where: TSN and MSN respectively refer to triply-supported and multiply-supported-networks. USN refers to the union of supported networks. Int, Tr, Ph refer to the Interactome, Transcriptome and Phenome networks respectively. Accuracy is the fraction of pairs gene linked in the network of interest, that is, gene pairs that share a specific GO term. Sensitivity is the fraction of gene pairs that share a specific GO term, which are linked in the network of interest.

	# of same function linked pairs	# linked pairs	accuracy(%)	sensitivity(%)	P-value
TSN	34	37	92	0.24	5e-37
MSN	595	680	88	4.1	<1e-300
USN	3945	11819	33	27	<1e-300
Int	272	353	78	1.9	7e-249
Tr	800	1382	58	5.6	<1e-300
Ph	3499	10801	32	24	<1e-300

Multiply-Supported Network

A portion of this integrated early embryogenesis network containing only links with two or more types of functional support was examined (Figure 19). In contrast to the full network, the topology of MSN con-

tains about half (305) of the early embryogenesis proteins-reveals distinct groups of highly interconnected genes/proteins and few or no links between the groups. Clusters were found using the greedy method of [1].

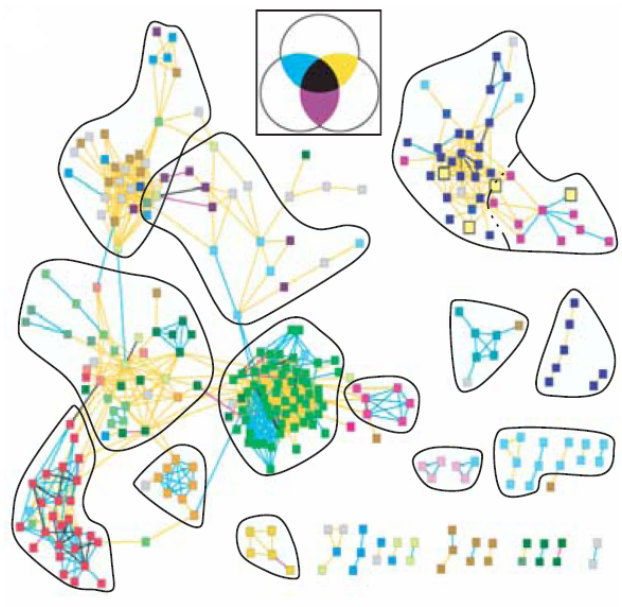


Figure 19: MSN containing 305 nodes joined by 1,036 edges, each supported by two or three types of functional evidence. Predicted molecular machines are encircled. Nodes are color-coded by function.

Results

Two types of highly interconnected regions were identified:

1. Modules containing a high density of links supported by both protein interactions and phenotypic correlations. These modules represent known molecular complexes that constitute discrete molecular machines within the cell. Virtually all of the edges in the graph that are supported by all three types of evidence (TSN) (41 out of 43 edges between 50 nodes) fall into such complexes. Proteins within such complexes function together as one physical unit, and depletion of any single member is likely to result in a very similar phenotypic profile.

2. Modules dominated by edges supported by both phenotypic and expression correlations, containing few physical interactions. These modules harbor genes that participate in distinct yet functionally interdependent cellular processes. Examples include messenger mRNA vs. protein metabolism. Within these models smaller molecular machines were found, supported by physical interactions and phenotypic similarity. Because current interactome maps have sampled only a small fraction of true interactions, such coordinated process modules may serve to predict undiscovered protein interactions. Alternatively, these modules may represent a qualitatively different type of functional unit, in which the phenotypic and expression profiling links reflect functional interdependencies dictated by the logical structure of the network, while the few protein interactions represent the physical path of information flow.

3 Pathway reconstruction

In this section we will describe two studies [9] [8] that aim at explaining knockout experiments.

3.1 Regulatory Pathway Reconstruction

In order to estimate which genes and proteins are regulated by a certain gene, the knockout method can be used. In this method the gene is inactivated. As a result, one can assess the differences in expression levels of other genes. For example, if a certain gene or protein was down- regulated by the knockout gene, its expression levels in the knockout sample should be higher than the one in the wild type sample, and vice-versa.The method presented in the works, tries to find the way in which a gene knockout affects other genes. The method looks at several knockouts and tries to explain the affects on genes along the path such that the path effects remain consistent to all knockouts. (Figure 20)

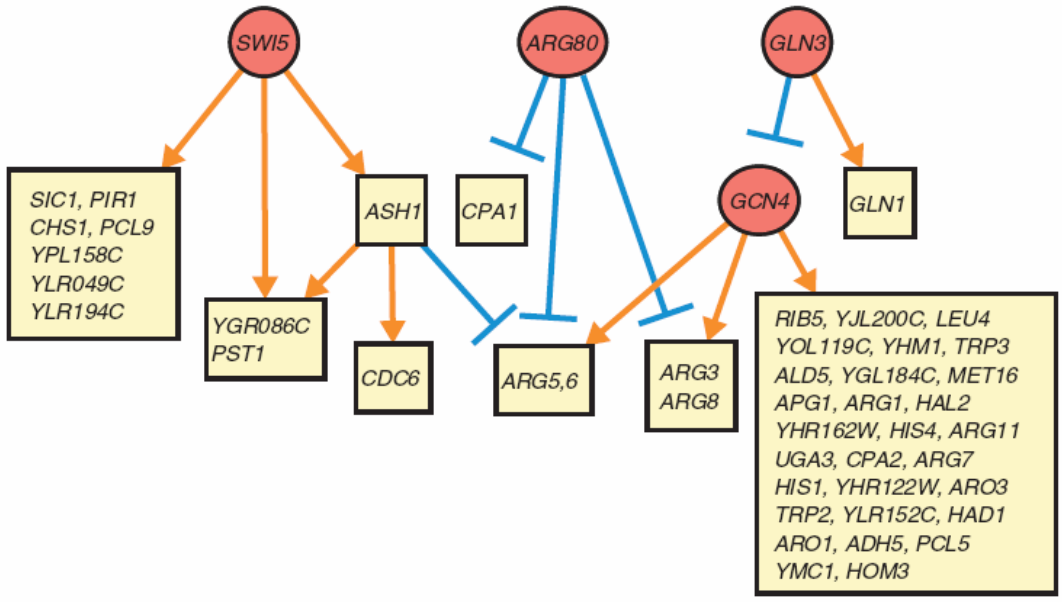


Figure 20: Effects of knockout of genes. The orange arrows show regulations and the blue arrows shows activation. If one of the regulators is inactivated it will cause the regulated gene to change its expression level.

3.2 Simplifying assumptions

Due to the high complexity of the regulatory system some simplifying assumptions need to be taken:

1. The effect is mediated by a linear PPI/TI pathway that is directed from the knockout to the affected gene and ends with a transcriptional edge. The pathway is assumed to have bounded length is necessary for computational reasons.

2. The effect propagates through the pathway (no combinatorial regulation involved).

3. Each edge has a sign: activation/repression. These should be consistent with the knockout effect (aggregate of signs should be opposite to the knockout effect). The sign is not known and our goal is

to assign each edge with a sign. We would like to find paths that are consistent with the assignments of signs to the edges. Note that there may be more than one way of assigning signs to edges (Figure 21)

4. Knockouts do not affect regulatory circuitry (the interaction data comes from wild-type conditions).

5. Assume physical interactions (PPI & TI) and effects are known (otherwise their noisy observations has to be modeled). This assumption should be dismissed when dealing with the same problem in research.

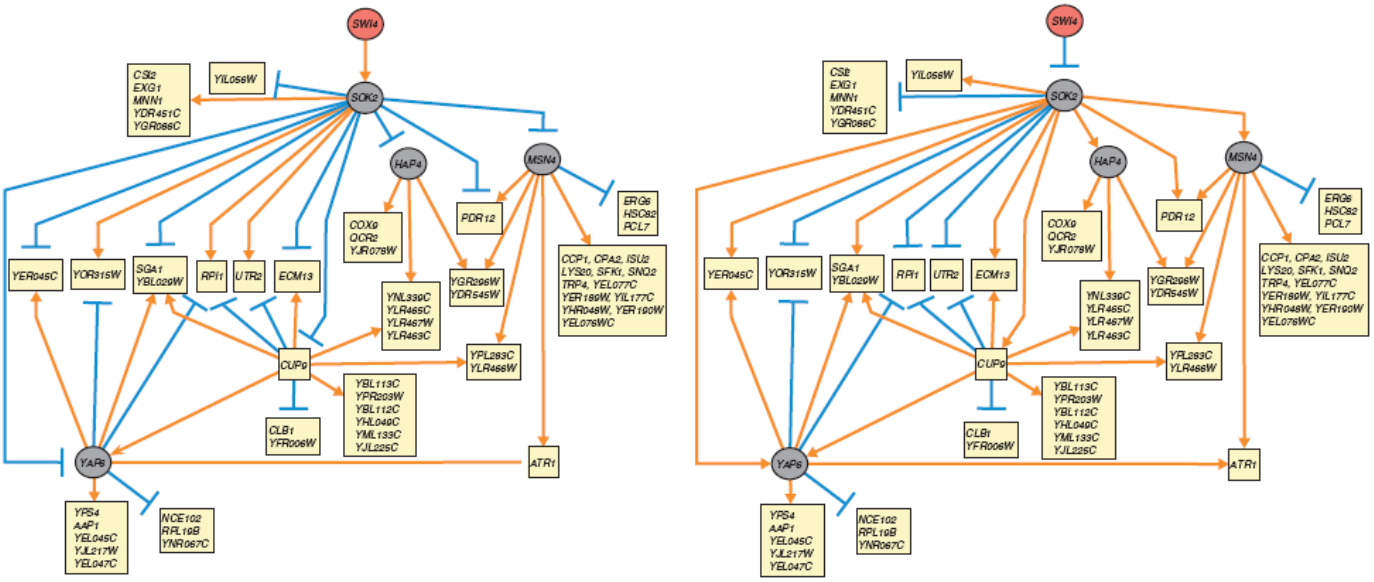


Figure 21: An example of different ways of explaining the same knockout effect from swi4 to msn4.

3.3 Probabilistic model

A probabilistic model is used to model the effect of knockout i on gene j via a product of potential functions. Let $\pi_{i,j}$be the set of paths from i to j that end with a transcriptional edge and obey the length constraint. For path a in $\pi_{i,j}$define a potential function

$$\psi_{i,j,a}(S_a, D_a, k_{i,j}, \sigma_{i,j,a}) = \epsilon_1 + (1-\epsilon_1) \cdot I(\sigma_{i,j,a} = 1) \cdot I(\prod_{e \in E_a} s_e = -k_{i,j}) \cdot \prod_{e \in E_a} I(d_e = \bar{d}_e)$$

where,

- ϵ_1- allow other causes for explanation, like inaccuracy in our models
- $I(\sigma_{i,j,a} = 1)$- path a explains the effect
- $I(\prod_{e \in E_a} s_e = -k_{i,j})$- signs on the path are consistent with effect
- $I(d_e = \bar{d}_e)$ - directions are consistent with the path

The probability function is a normalized product of the potential functions.
Forcing at least one explanatory path implies:

$$\psi \underset{i,j}{OR} (\sigma_{i,j,1}...\sigma_{i,j,|\prod_{i,j}|}) = \epsilon + (1-\epsilon) \cdot (1 - \prod_a I(\sigma_{i,j,a} = 0))$$

The complete potential function for (i,j) effect is:

$$\psi_{i,j}(S_{i,j}, D_{i,j}, \Sigma_{i,j}, k_{i,j}) = \psi \underset{i,j}{OR} (\sigma_{i,j,1}...\sigma_{i,j,|\prod_{i,j}|}) \cdot \prod_a \psi_{i,j,a}(S_a, D_a, k_{i,j}, \sigma_{i,j,a})$$

3.4 Experimental Validation

Mating response pathways were analyzed. For 149 effects, under 13 deletion experiments, 106 were connected via candidate paths in the network, all of which were explained by the model. In order to validate the results, cross validation was used.

Cross-validation is the statistical practice of partitioning a sample of data into subsets such that the analysis is initially performed on a single subset, while the other subsets are retained for subsequent use in confirming and validating the initial analysis. The initial subset of data is called the training set; the other subsets are called validation or testing sets. Holdout validation is not cross-validation in common sense, because the data never are crossed over. Observations are chosen randomly from the initial sample to form the validation data, and the remaining observations are retained as the training data. Normally, less than a third of the initial sample is used for validation data. In this case, assuming a path with assigned signs, when trying to knockout a gene in the middle of the path, it confirms the effect as predicted. The following table shows the error rate as for different held-outs:

#hold-outs	#trials	%error
1	106	2.83%
5	200	3.5%
20	200	5.9%

The authors also provided a sensitivity analysis to exclude the possibility that the results may have been artifacts of a particular setting of the model parameters/thresholds. They considered the following adjustable parameters: the maximum length of candidate paths, thresholds on p-values of selecting candidate protein-DNA and knock-out pairs, and the error probabilities used as soft constraints in the potential functions. Figure 22 examines the sensitivity of the model for various thresholds.

3.5 Genome-wide Application

The method was tested globally on genome-wide including data of yeast, 5500 TIs, 15000 PPIs and 273 deletion profiles. Regulatory effects were uniquely determined for a small fraction of the participating interactions: 194/771 (part of it can be seen in Figure 23).

After applying the method there are still about 500 cases that cannot be uniquely explained well enough. An attempt was done to use the method to decide what kind of experiments should be done. Next the data was fragmented and 37 distinct regions remained ambiguous. 20 of these regions correspond to known pathways or are functionally enriched. Figure 18 shows one region of the 37 where the paths remained ambiguous. The goal is to design new experiments that will resolve maximum number of ambiguities in order to model the network. Achieving this goal will allow to solve the ambiguities in minimum number of experiments. The design used the following method: Rank deletion experiments by their expected information gain. The expected information was calculated using

$$I(M; Y^e) = H(M) - H(M|Y^e) = H(M) + \Sigma_{m,y} P(M = m, Y^e = y) \log_2 P(M = m|Y^e = y)$$

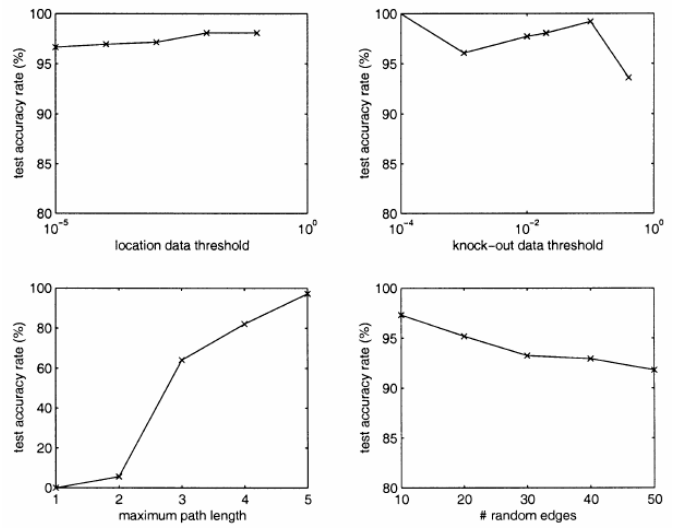


Figure 22: The accuracy as a function of different thresholds.

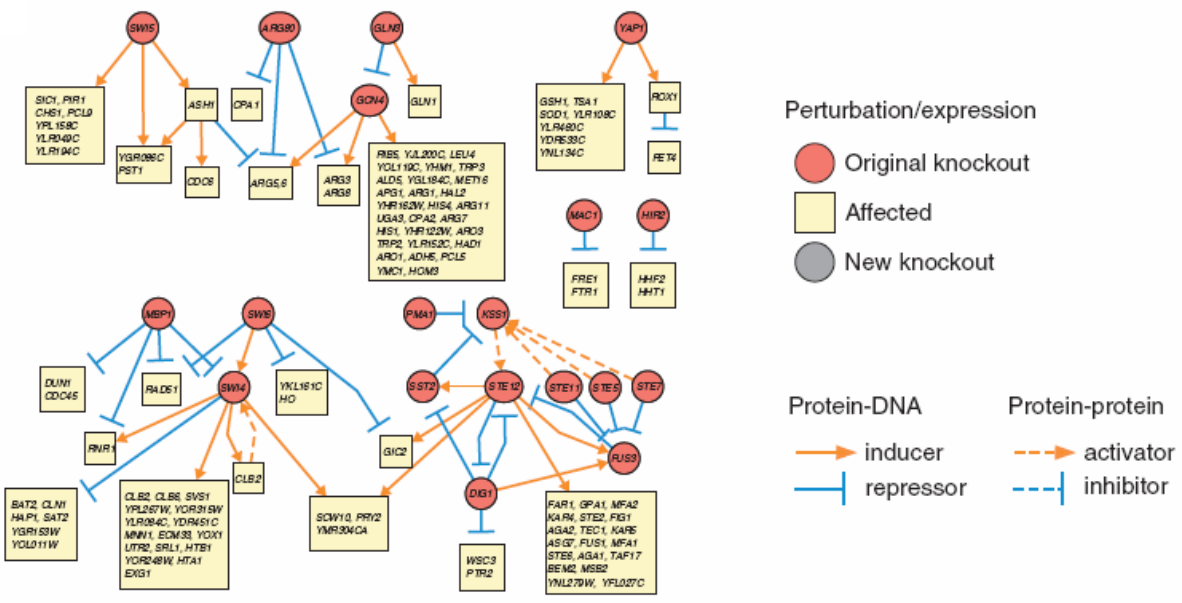


Figure 23: Fraction of the network where the regulatory effects were uniquely determined.

where M is set of possible model configurations, and Y^e is the predicted effects for experiment e.

The four top-scoring experiments were performed (deletions for sok2, yap6, hap4, msn4). As a result from these experiments 60 TIs were disambiguated.

4 Network dynamics

All the works that were described in previous chapters considered a static model of a network, examining all the connections in those networks. A question that needs to be asked is what if the connections in the network are related to different conditions? The work of [5] considers a dynamic network model in which the connections depend on the condition the system is currently in.

4.1 Condition-Specific Networks

Physical network provides a static picture. However, not all interactions are active in a given condition/time point. Dynamic information can be obtained by combining condition-specific data such as gene expression. Here TIs are combined with gene expression profiles in five conditions: cell cycle, sporulation, diauxic shift, DNA damage and stress response. Figure 24 shows the static network. Figure 25 shows the activation of genes under different conditions. Half of targets are uniquely expressed in one condition. Half of the interactions change between conditions (between every two situations). 66 interactions are active over ≥ 4 conditions, mainly regulating house-keeping genes, that are necessary for the cell.

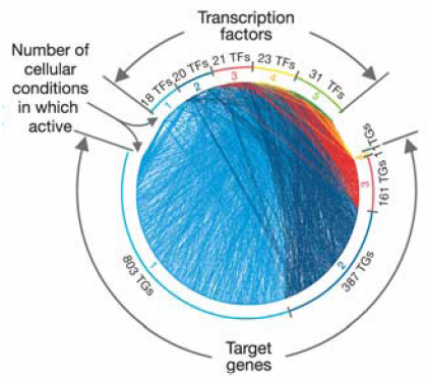


Figure 24: Graphs of the static and condition-specific networks. Transcription factors and target genes are shown as nodes in the upper and lower sections of each graph respectively, and regulatory interactions are drawn as edges; they are colored by the number of conditions in which they are active. Different conditions use distinct sections of the network.

4.2 Trace back Algorithm

In order to decide if a gene is active under a certain condition the trace back algorithm was used. It is a naive algorithm which determines when a gene is active. The algorithm works as follows: It identifies transcription factors as being 'present' in a condition if they have sufficiently high expression levels. It then marks as 'active' the regulatory links between present transcription factors and differentially expressed genes; finally it searches for any other present transcription factors that are linked to a transcription factor with an already active link and makes this connection active. The last step is repeated until no more links become active.

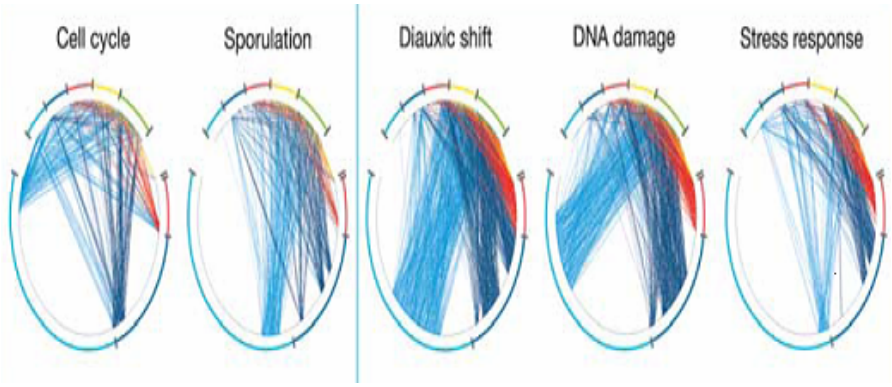


Figure 25: The network in different stages. Again, transcription factors and target genes are shown as nodes in the upper and lower sections of each graph respectively, and regulatory interactions are drawn as edges.

4.3 Endogenous vs. Exogenous Systems

The conditions that the network was examined under could be categorized into two groups with different biological traits: endogenous and exogenous. This allows to rationalize the different sub network structures discovered in terms of the biological requirements of each condition.

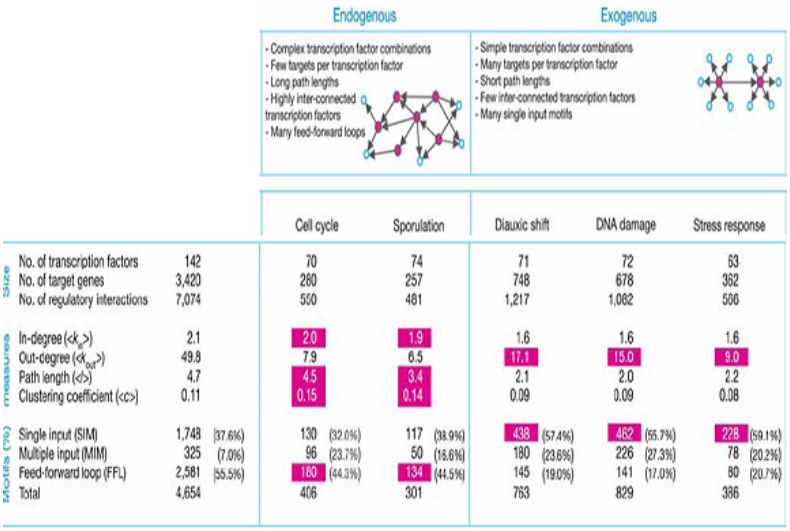
- Endogenous Networks are multi-stage processes, operating with an internal transcription program. These are coordinated processes with long paths (intermediate phases), high clustering (inter-regulation between TFs) and large in-degree (complex multiple-TF regulation).
- Exogenous Networks constitute binary events that react to external stimuli with a rapid turnover of expressed genes. The rapid response are along short paths (faster signal propagation) and large out-degrees (each TF regulates many genes).

Figure 26 shows the info that was hidden in static networks and could be found only when looking an the dynamic network. The shadowed values in purple represent significant values.

4.4 Permanent vs. Transient Hubs

In response to diverse stimuli, transcription factors alter their interactions to varying degrees, thereby rewiring the network. A few transcription factors serve as permanent hubs, whereas most act transiently only during certain conditions [5]. Hubs are of general interest as they represent the most influential components of a network and, accordingly, tend to be essential. They are thought to target a broad spectrum of gene functions and are commonly located upstream in the network to expand their influence via secondary transcription factors. These observations suggest that hubs would be invariant features of the network across conditions, and this expectation is supported by the random simulations that converge on similar sets of transcription factor hubs.

When checking the hubs under different conditions it was found that 78% of the hubs were influential in only one condition. That means that most of the hubs in the network are condition specific (as can be seen in Figure 27). Exogenous conditions have fewer hubs, suggesting a more centralized command structure. About half of the transient hubs are known to be important for their respective conditions. For the remainder with sparse annotations, their transient-hub status in a particular condition considerably augments their functional annotation. The defining feature of transient hubs is their capacity to change interactions between conditions.



		Cell cycle	Sporulation	Diauxic shift	DNA damage	Stress response
Size	No. of transcription factors	142	70	74	71	72	63
	No. of target genes	3,420	260	257	748	678	362
	No. of regulatory interactions	7,074	550	481	1,217	1,062	506
measures	In-degree ($\langle k_{in} \rangle$)	2.1	2.0	1.9	1.6	1.6	1.6
	Out-degree ($\langle k_{out} \rangle$)	49.8	7.9	6.5	17.1	15.0	9.0
	Path length ($\langle d \rangle$)	4.7	4.5	3.4	2.1	2.0	2.2
	Clustering coefficient ($\langle c \rangle$)	0.11	0.15	0.14	0.09	0.09	0.08
Motifs (%)	Single input (SIM)	1,748 (37.6%)	130 (32.0%)	117 (38.9%)	438 (57.4%)	462 (55.7%)	228 (59.1%)
	Multiple input (MIM)	325 (7.0%)	96 (23.7%)	50 (16.6%)	180 (23.6%)	226 (27.3%)	78 (20.2%)
	Feed-forward loop (FFL)	2,581 (55.5%)	180 (44.3%)	134 (44.5%)	145 (19.0%)	141 (17.0%)	80 (20.7%)
	Total	4,654	406	301	763	829	386

Figure 26: Network statistics across different conditions. Standard statistics (global topological measures and local network motifs) describing network structures. These vary between endogenous and exogenous conditions; those that are high compared with other conditions are shaded. (Note, the graph for the static state displays only sections that are active in at least one condition, but the table provides statistics for the entire network including inactive regions.)

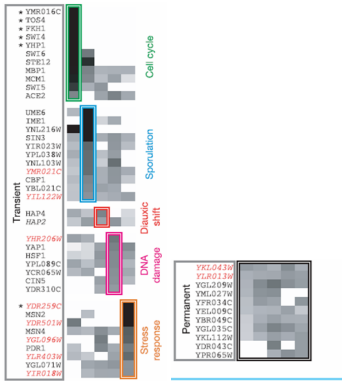


Figure 27: Hubs active in different situations.

4.5 Phase-specific TFs

The article [5] also tries to observe the dynamics within a process, that is, to see how the gene corresponds to different phases in the same process. The motivation of examining this comes because some of the conditions change across time (like the cell cycle). Testing the different phases of the cell cycle will allow better understanding of them. When looking at the cell cycle it is apparent that most of the TFs are phase specific (as can be seen in Figure 28). The TFs regulate each other in a serial manner, where every TF regulates the TFs in the next phase.

4.6 Interaction Interchange

Figure 29 shows that most hubs change 10-90% of their interactions between conditions. The authors of the article described above attempted to quantify this rewiring more broadly for every transcription factor in the network with the interchange index, I. This is defined so that higher values associate with transcription factors replacing a larger fraction of their interactions. At one extreme ($I \leq 10\%$), 12 transcription factors retain all interactions across multiple states. At the other end ($I \geq 90\%$), 27 transcription factors replace all interactions in switching conditions. Many of these are so extreme that they only regulate genes in a single condition and are inactive otherwise. Most transcription factors interchange only part of their interactions ($10\% < I < 90\%$). This group comprises most of the hubs; surprisingly, permanent hubs interchange interactions as often as transient ones, but over a larger number of conditions. Furthermore, transcription factors in this group often regulate genes of distinct functions in different conditions, thus shifting regulatory roles.

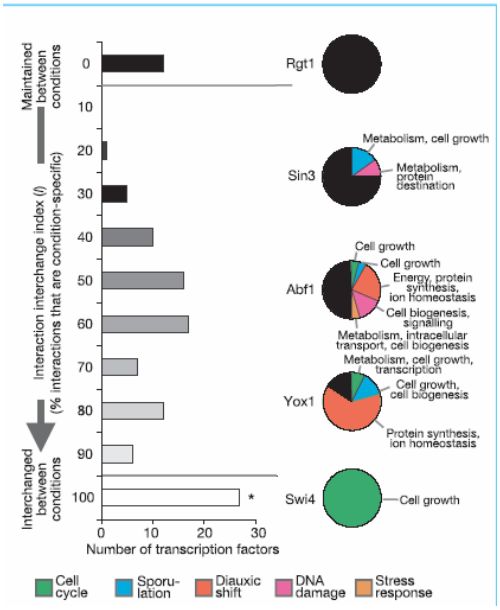


Figure 28: Interaction interchange (I)of transcription factors between conditions. A histogram of I for all active transcription factors shows a uni-modal distribution with two extremes. Pie charts show five example transcription factors with different proportions of interchanged interactions. The main functions of the distinct target genes regulated by each example transcription factor were listed. Note how the transcription factors regulatory functions change between conditions.

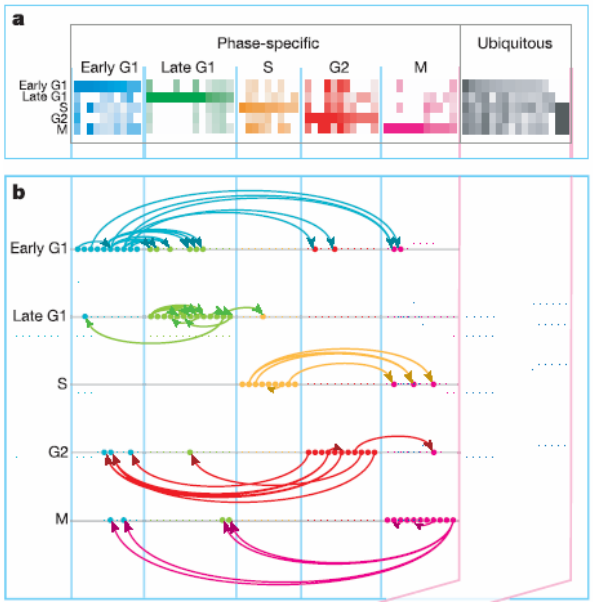


Figure 29: Transcription factor inter-regulation during the cell cycle time-course. a, The 70 transcription factors active in the cell cycle. The diagram shades each cell by the normalized number of genes targeted by each transcription factor in a phase. Five clusters represent phase-specific transcription factors and one cluster is for ubiquitously active transcription factors. Note, both hub and non-hub transcription factors are included. Transcription factor names are given in the Supplementary Information. b, Serial inter-regulation between phase-specific transcription factors. Network diagrams show transcription factors that are active in one phase regulate transcription factors in subsequent phases. In the late phases, transcription factors apparently regulate those in the next cycle.

5 Summary

We could see many advantages for integrating networks: Different networks provide different views of cellular processes and are inter-related. Data integration reinforces functional modules supported by several data sources. It allows more accurate models of biological processes and their reconstruction, and also allows elucidation of network dynamics. Finally, it improves predictions of protein function and interaction.

References

[1] G. D. Bader and C. W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(2), 2003.

[2] Milo et al. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[3] Gunsalus et al. Network predictive models of molecular machines involved in caenorhabditis elegans early embryogenesis. *Nature*, 436(7052):861–865, 2005.

[4] R. Kelly and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23:561–566, 2005.

[5] Luscombe et al. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431:308–312, 2004.

[6] S. Shen-Orr et al. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.

[7] K. Tan et al. Transcriptional regulation of protein comlexes within and across species. *PNAS*, 104(4):1283–1288, 2007.

[8] Yeang et al. Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biology*, 6(7), 2005.

[9] CH. Yeang, T. Ideker, and T. Jaakkola. Physical network models. *Journal of computational biology*, 11(2-3):243–262, 2004.

[10] E. Yeger-Lotem et al. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, 101(16):5934–5939, 2004.

[11] Zhang et al. Motifs, themes and thematic maps of an integrated saccharomyces cerevisiae interaction network. *J.Biol*, 4(6), 2005.