BIOINFORMATICS

Vol. 18 Suppl. 2 2002
Pages S202–S210

Building and analysing genome-wide gene disruption networks

J. Rung [1,†], T. Schlitt [1,†], A. Brazma [1], K. Freivalds [2] and J. Vilo [1]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [2] Institute of Mathematics and Computer Science, University of Latvia, Riga, LV - 1459, Latvia

Received on April 8, 2002; accepted on June 15, 2002

ABSTRACT

Motivation: Microarray experiments comparing expression levels of all genes in yeast for hundreds of mutants allow us to examine properties of gene regulatory networks on a genomic scale. We can investigate questions such as network modularity, connectivity, and look for genes with particular roles in the network structure.

Results: We have built genome-wide disruption networks for yeast, using a representation of gene expression data as directed labelled graphs. Nodes represent genes and arcs connect nodes if the disruption of the source gene significantly alters the expression of the target gene. We are interested in features of the resulting disruption networks that are robust over a range of significance cutoffs. The networks show a significant overlap with analogous networks constructed from scientific literature. In disruption networks the number of arcs adjacent to different nodes are distributed roughly according to a power-law, like in many complex systems where the robustness against perturbations is important. The networks are dominated by a single large component and do not have an obvious modular structure. Genes with the highest outdegrees often encode proteins with regulatory functions, whereas genes with the highest indegrees are predominantly involved in metabolism. The local structure of the networks is meaningful, genes involved in the same cellular processes are close together in the network.

Keywords: gene networks, microarrays, yeast, graph visualisation

Availability: http://www.ebi.ac.uk/microarray/networks

Contact: schlitt@ebi.ac.uk; johan@ebi.ac.uk

INTRODUCTION

By measuring mRNA expression levels of thousands of genes in parallel, microarray experiments help revealing the structure of the underlying gene regulatory networks. The cellular regulatory system is a complex mechanism, and there is no straightforward way to represent it as a simple graphical model. The reduction of information needed to create a graphical model with nodes and arcs can be done in many ways, resulting in a wide range of network models with different interpretations.

A *gene network* is a directed labelled graph, where each node represents a gene and each arc represents a relation between the genes. The direction and the labels attached to an arc represent the nature and strength of the relation or the evidence for it. For example, an arc can mean that the source gene is coding for a transcription factor known to be binding to the promoter of the target gene. We obtain a different network if we define an arc as an observation in gene expression data, that the change in the expression level in the the source gene implies the change in the expression in the target gene. A network can also be built from literature data, linking genes which have been mentioned in the same paper (Stapley and Benoit, 2000; Jenssen *et al.*, 2001).

Various methods have been used to build gene expression networks, for instance, Bayesian networks (BN) (Friedman *et al.*, 2000) and Dynamic Bayesian networks (DBN) (Murphy and Mian, 1999). Pe’er *et al.* describe a method to infer subnetworks of interacting genes from gene expression data in a BN framework (Pe’er *et al.*, 2001). They are interested in finding features with high confidence in BNs learnt from a set of 565 genes in mutation experiments for *Saccharomyces cerevisiae* (Hughes *et al.*, 2000). Because the data set is too small to yield a single high confidence network, they use bootstrapping techniques to find such features. These are extracted into subnetworks, which are interpreted as separate cellular processes or putative interactions.

Although BN based methods are powerful techniques, the approaches that have been implemented currently are only able to deal with relatively small subsets of genome data. We have chosen an approach that allows for a genome-wide analysis, and although simple, we can demonstrate that it is biologically meaningful and provide insight in the genome-wide organisation of the gene networks. We build gene disruption networks by

[†] These authors contributed equally to this paper.

collecting information about differences in gene expression between yeast mutant strains. Nodes represent genes and arcs connect the deleted genes with the genes for which expression level changes have been observed for a particular threshold. We compare these networks to a network derived from literature data and explore their properties, including the distribution of arcs and the cellular functions of genes with many interactions as well as the topology and robustness of the network for different discretisation thresholds. Finally, we present an example of a subnetwork including genes involved in the pheromone response pathway.

BUILDING THE GENE DISRUPTION NETWORKS

A *directed graph* is defined as a tuplet $(\mathcal{G}, \mathcal{A})$ of nodes $\mathcal{G}$ and arcs $\mathcal{A}$, where an arc $a \in \mathcal{A}$ is an ordered pair of nodes $(g_1, g_2) \subseteq \mathcal{G}$. We can attach labels to nodes and arcs, in which case we obtain a *labeled graph*. In a *gene disruption network* Δ, each node represents a gene. A gene g_1 is connected to gene g_2 with an arc $a = (g_1, g_2)$, if the disruption of the gene g_1 changes the expression level of gene g_2 significantly. The ‘significance’ is defined through a threshold as described below.

The starting point for our network building is a *gene expression data matrix* E, in which each experimental condition, which in our case is a disruption of a particular gene, corresponds to a column and each gene corresponds to a row. The jth element in the ith row holds the ‘expression level’ r_{ij} of gene i in experiment j, more precisely $r_{ij} = \log(l_{ij}/c_{ij})$, where l_{ij} is the background corrected signal for the studied condition relative to that of the same gene in a control c_{ij}.

Next, we transform the original gene expression data matrix E into a discretised matrix D, where values $d_{ij} \in \{-1, 0, 1\}$ represent the expression level being decreased, unchanged or increased with respect to the control experiment. For this we first normalise the log-ratios r_{ij} to $\tilde{r}_{ij} = r_{ij}/\hat{\sigma}_{ij}$, where $\hat{\sigma}_{ij}$ is a gene-specific standard deviation estimate for the measurement of gene i under experimental condition j. This can be estimated using different error models, see for instance (Hughes *et al.*, 2000). The discretisation is carried out on the normalised data using a cutoff level $\gamma > 0$, and defining

$$d_{ij} = \begin{cases} -1, & \tilde{r}_{ij} \leq -\gamma \\ 0, & -\gamma < \tilde{r}_{ij} < \gamma \\ 1, & \tilde{r}_{ij} \geq \gamma. \end{cases} \tag{1}$$

Due to the gene-specific normalisation, we can use one consistent cutoff level for all measurements.

Effectively the disruption network $\Delta(\gamma) = (\mathcal{G}, \mathcal{A})$ is a representation of the discretised matrix as a graph. For the given cutoff γ, we draw an arc from gene $g_j \in \mathcal{G}$ to gene $g_i \in \mathcal{G}$, if $d_{ij} \neq 0$. We can label the arcs as downregulating or upregulating, depending on whether $d_{ij} = -1$ or $d_{ij} = +1$, respectively. In the graphical representation we can show this by drawing them in red and green (or solid and gray in black-and-white representation). We also label the nodes by their respective gene names. Examples of gene disruption networks are given in Figure 4 and Figure 5.

We use the microarray data set from Hughes *et al.* (Hughes *et al.*, 2000), which includes expression profiles for all genes in *Saccharomyces cerevisiae* over a set of 300 experiments. In 274 experiments single gene deletion mutants were examined, 2 experiments were double gene deletion mutants, 13 experiments were done with genes with tetracyclin regulated promoters and in the remaining 11 experiments the yeast cell cultures were treated with different drug compounds.

In parallel to these experiments, a series of 63 control experiments were performed, comparing untreated wild-type yeast cultures to each other, permitting the use of a gene specific error model and to normalise the data using the standard deviation estimates. Following normalisation we discretised the data matrix using different significance cut-offs between $\gamma = 1.0$ and $\gamma = 26.0$, which was the threshold where no edges at all were found. However, the analyses described below concentrate on the range $\gamma = 2.0, 2.1, \ldots, 4.0$. We included only genes with less than 25% undefined data points in the gene expression matrix as given in (Hughes *et al.*, 2000).

NETWORK VISUALISATION

The analysis of the network properties is facilitated by their visualisation based on a graphical layout in 2 dimensions. The gene disruption networks may be very large and dense, therefore their visualisation is not trivial. To visualize the obtained networks we used the Graphical Diagramming Engine (GDE) (Freivalds and Kikusts, 2001). GDE supports five layout styles, each revealing a different aspect of the network. In the hierarchical style drawings it is easy to notice the nodes with high in- and out-degree and their relationships. The spring-embedder layout reveals the cluster structure of the network, clearly identifying the parts that are strongly related. The grid layout was used to produce the drawings in this paper; it gives the most compact drawings useful when the area for displaying the drawing is limited. The tool also provides several edge representations including orthogonal, polyline and spline. Here the spline representation was the most adequate since the smoothness allows easier following of the edge. Although layouts of these networks were generated automatically, the rich editing facilities powered by the quadratic optimization method (Freivalds and Kikusts, 2001) were used in exploring the networks, discovering their features and preparing their relevant parts for presentation. For connected component placement an original polyomino packing algorithm (Freivalds *et al.*, 2001) was used. Such an approach allows a compact placement and high visual separation of the components.

COMPARISON OF DISRUPTION NETWORKS WITH A LITERATURE NETWORK

In order to relate the disruption networks to biological results, we compared them to a reference network constructed using the curated database YPD (Costanzo *et al.*, 2001), which contains information about the yeast genome.

A *YPD network* is a graph, where the nodes represent genes and two genes are connected with an undirected edge,if gene X is mentioned in the description of gene Y in the YPD database. Thus an edge between X and Y could be read as 'the database entry for X mentions Y and/or vice versa'. Note that we have not done any further analysis of the syntax, e.g. if the original description was 'a change in x does not influence y' we would still have an edge. These cases are rare and do not diminish the overall use of the reference network. The lack of text analysis is also the reason why we use undirected edges instead of directed edges, e.g. if the description for gene X reads 'X regulates Y' then there is also a relation 'Y is regulated by X'.

For our study, we selected a subset of 274 genes studied in (Hughes *et al.*, 2000). The resulting network has 827 edges. If our disruption networks represent meaningful biological knowledge, we would expect that it should overlap with the YPD network more than expected by chance. We checked this assumption by calculating the overlap between the disruption networks and the reference network, then comparing it to the distribution of overlaps between randomised networks and the reference network. Each randomised instance of the disruption network was built by keeping the number of arcs present in the disruption network constant. Every node has on average the same in- and outdegree as in the original network. This procedure assures that the overall network topology is retained.

We created 1000 randomised networks for each disruption network built. All disruption networks showed a significantly higher overlap with the reference network than the corresponding randomised networks (see Figure 1). In the range of $\gamma \in \{2.0, 2.1, \ldots, 4.0\}$, the overlap between the disruption networks and the reference network inceases monotonously from 9–16%. If we interpret the reference network as representing connections between genes reported in literature, this result demonstrates that the disruption networks contain significantly more biological information than achieved randomly.

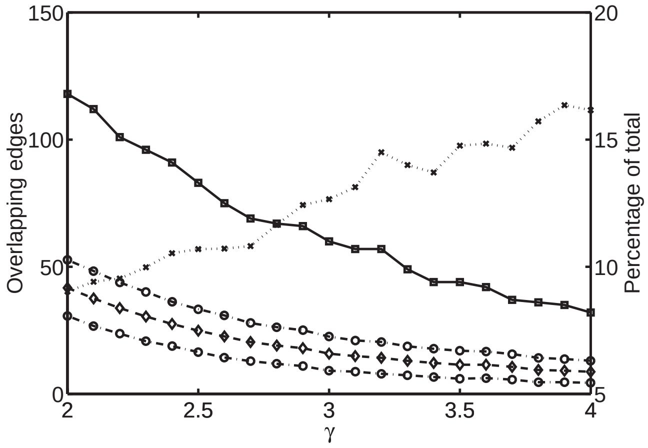

Fig. 1. Disruption networks compared with randomised ones using a reference network based on YPD data. The solid line shows the number of arcs which are found in both the disruption network for cutoff level γ and the YPD network. The dashed line shows the average overlap between 1000 randomised networks and the same YPD network, with 2 standard deviations shown by the dash-dotted line. The dotted line shows the percentage of edges in the disruption networks that are also present in the YPD network. For the range of γ studied, this statistic is monotonously increasing and no obvious optimisation can be done based on it.

'IMPORTANT' GENES AND GENES WITH COMPLEX REGULATION

The *degree of a node in a graph* is defined as the number of adjacent edges. In directed graphs we distinguish between the *indegree*—the number of incoming arcs—and the *outdegree*—the number of outgoing arcs. For disruption networks we can speculate that genes with a high outdegree are 'important' in the sense that they influence the expression of many other genes, while genes with a high indegree have a complex regulation mechanism. In order to analyze the indegree and outdegree of various genes we represent them in two different formats: the *degree table*, where the genes are sorted according to their outdegree or indegree, together with their annotation in the YPD database (Costanzo *et al.*, 2001); the *cellular role table*, where we use the 'cellular role' annotation from YPD to group all genes with at least one arc in a particular network and calculate for each group the median indegree and outdegree, and sort the groups according to their median degree.

We find that the distribution of total degree, the sum of the in- and out-degree for each node, roughly follows a power-law (Figure 2). This topology, denoted *scale-free* (Barabasi and Albert, 1999), has been found for metabolic networks (Jeong *et al.*, 2000) and for many other complex systems (Albert and Barabasi, 2002).

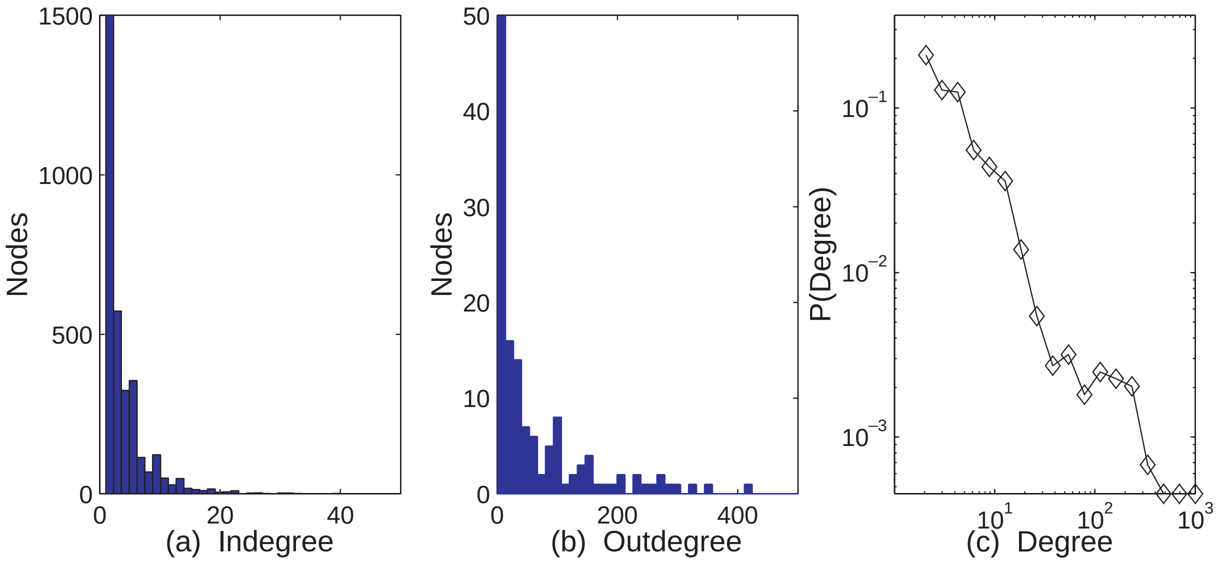

Fig. 2. (a,b) Histogram of number of nodes (a) with particular indegree or outdegree (b) for $\gamma = 2.5$. (c) Log–log plot of the distribution of total degree for the disruption network with $\gamma = 2.5$

The genes with the highest outdegrees encode proteins involved in transcriptional regulation (TUP1, SHE4, SWI4, CYC8)[†], a ribosomal protein (RPL12A), a protein involved in rRNA modification and degradation of aberant mRNA (RRP6), a histone deactylase (RPD3), a MAP kinase (KSS1), proteins involved in metabolism (ANP1, ERG2, FKS1) and a H^+-ATPsynthase subunit (CUP5). It is quite remarkable that in the case of the TUP1 deletion mutant about 50% of the genes show changes in expression and still the yeast cells survive. A full degree table for $\gamma = 2.0$ is included in our supplementary data.

The order of the groups in the cellular role table is robust over a wide range of cut-off values (see Table 1), though the number of genes per group is relatively small since only 248 mutants were included in the data set. The cellular roles with the highest median outdegrees predominantly have regulatory functions. It has to be kept in mind that genes can belong to more than one of these groups and sometimes the annotation can be misleading. RTG1 for example has an outdegree of 316, it is a transcription factor, but it is also a member of the ‘carbohydrate metabolism’ group.

The overwhelming majority of genes with the highest indegree are involved in metabolism (ADE17, VID24), especially amino acid biochemistry (YGL009C, HIS5, HIS1, TWT1, HOM3), stress response (HSP12, YHB1), transport (SIT1, PEX21, ARN1, YHM1) as well as some genes of unknown function (see supplementary data). The ordering of the groups in the cellular role table for the indegree is robust over a wide range of cut-off values. The group sizes are much larger compared to the cellular role table for the outdegree (see Table 1), therefore the grouping is more reliable.

Only a few genes can be found to have a high indegree and a high outdegree at the same time. When we plot the genes using their rank for outdegree and indegree in the degree table we find that ARG5,6 with an outdegree of 108 and an indegree of 28 (see Figure 3) is the only gene which is within the top 50% of the genes with highest outdegree and highest indegree.

The comparison of the cellular roles of the genes with the highest out- and in-degrees in the yeast mutation networks seem to confirm the intuition that general regulators are influencing many genes, whereas some metabolic genes are being regulated by many other genes. Note that the only group having a high median indegree and outdegree is ‘small molecule transport’.

CONNECTED COMPONENTS

Recently, some investigations on the overall structure of gene networks have been published. Wagner predicts the existence many independent subnetworks and only a few direct connections for each gene (Wagner, 2002), whereas Featherstone and Broadie argue that there is ‘a single giant functional component rather than several subnetworks’. They also found that hubs, the genes with highest degrees, are evolutionary more conserved than the genes with lower degrees. (Featherstone, 2002).

[†] We use capital letters to refer to genes in the disruption networks.

Table 1. Cellular role table showing the top 5 groups with the highest median degrees for selected networks, with a minimum group size of 3 for outdegree and 40 for indegree (m: the median degree, n: the group size)

γ	outdegree	m	n	indegree	m	n
2.0	carbohydrate metabolism	363	4	amino-acid metabolism	9	194
	RNA turnover	353	4	nucleotide metabolism	6	82
	meiosis	244	3	energy generation	5	242
	cell stress	207	9	*small molecule transport*	5	343
	protein translocation	197	3	other metabolism	5	148
2.8	RNA turnover	110	4	amino-acid metabolism	4	167
	cell stress	62	8	nucleotide metabolism	3	67
	meiosis	54	3	energy generation	2	184
	protein synthesis	53	7	differentiation	2	43
	cell wall maintenance	47	6	*small molecule transport*	2	286
3.6	RNA turnover	48	4	*small molecule transport*	2	230
	RNA processing/ modification	41	4	other metabolism	2	96
	cell stress	27	8	nucleotide metabolism	2	58
	small molecule transport	19	8	mating response	2	57
	cell wall maintenance	19	6	amino-acid metabolism	2	133

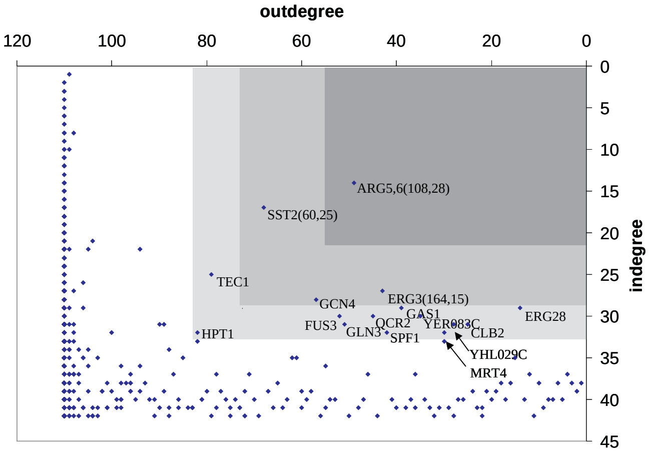

Fig. 3. Genes plotted according to their rank in the indegree and outdegree table for the network with $\gamma = 2.0$. Some outdegrees and the indegrees are given in brackets. The shaded areas indicate regions of ranking, darkest grey top 50%, medium grey top 66%, light grey top 75%. The only gene with a in- and out-degree within the top 50% for both is ARG5,6, a acetylglutamate kinase and N-acetyl-gamma-glutamyl-phosphate reductase.

To address the open question about modularity, we are interested in the connectivity structure of our networks and want to examine whether or not, and to what extent yeast disruption networks have a modular structure. A *connected component* in a graph is defined as a subgraph where there exists a path through arcs (ignoring the direction) leading from one node to the other for each pair of nodes. A single gene disconnected from all other genes is not regarded as a separate connected component. For a low enough threshold γ all genes will be connected.

For $\gamma \leq 3.0$ only one connected component can be found[‡]. For higher thresholds γ we find one dominant and few small components of sizes 2 or 3 genes. The dominant connected component consists of 5383 genes for $\gamma = 2.0$ and of 2354 genes for $\gamma = 4.0$ (see table 2).

Since the networks contain only a small number of nodes with rather high degrees, it is interesting to see if we can break down the connected component by removing the genes with the highest indegree or outdegree from the networks. For this we removed the top 1, 5 and 10% of genes when ranked for the outdegree, respectively indegree. This disrupts the connected component considerably. However, for $\gamma \leq 3.6$ we still find only one major component and some much smaller components even when removing the top 10% of the genes with highest degrees (see Table 2). The dominant components are reduced in size, but are still at least one order of magnitude bigger than the minor components.

Only the networks with $\gamma \geq 3.7$ consist of components of roughly equal sizes when the 10% of the genes with highest degrees are removed. Removing the top ten percent of the genes with the highest degrees at $\gamma = 3.7$ yields a network with 378 genes with at least one adjacent arc and a total of 331 arcs. The five biggest components have size 93, 53, 31, 20, 10 and there is a total of 50 components. Although in some of these components genes belonging to the same YPD annotation group are overrepresented, this is not true for the majority of the

[‡] There is an exception for the network at $\gamma = 2.6$, which has an additional component consisting of 2 nodes

Table 2. Size of the biggest two and the total number of connected components in selected networks, the column 'full' refers to the original network, the numbers in the other columns result from removing the genes with the highest degrees (1, 5, 10%: proportion of genes which were removed)

γ		full	removed		
			1%	5%	10%
2.0	biggest	5383	4707	3682	2614
	second			2	5
	total number	1	1	2	2
3.0	biggest	3556	2461	1385	764
	second	2	2	4	6
	total number	2	2	9	17
3.6	biggest	2789	1612	901	485
	second	2	3	10	11
	total number	3	4	13	26
3.7	biggest	2675	1497	825	93
	second	2	3	9	53
	total number	3	4	14	50
4.0	biggest	2354	1205	542	45
	second	3	3	6	28
	total number	4	7	22	51

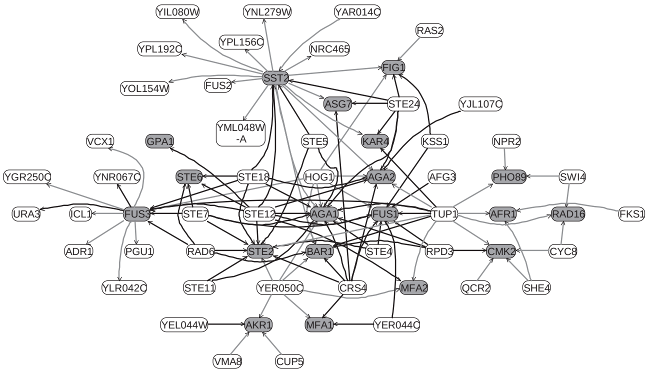

Fig. 4. This subnetwork is the result of filtering the full network at $\gamma = 4.0$ for the core set marked in grey and their next neighbours (grey arcs: downregulation, dark arcs: upregulation). See web supplement for a coloured figure (http://www.ebi.ac.uk/microarray/networks)

groups. It seems possible that these components may not have a true biological meaning, but are rather the result of the mechanical reduction of the graph.

Modules are thought to be building blocks which confer a particular function and might be interlinked to other modules via hubs. However, the results of Featherstone *et al.* and our analysis would argue for a closely connected network organisation rather than a modular structure. Genes involved in the same cellular processes seem to be closer in the network than unrelated genes (see also next section), however, this does not mean that the network can be easily broken down into independent parts. This indicates a scale free structure, which has the property of self-similarity, meaning that any part of the network is statistically similar to the whole network (Wolf *et al.*, 2002).

SUBNETWORKS AND THEIR VISUALISATION

Finally, we focus on subnetworks containing genes of particular interest, as an example taking the pheromone response. When yeast cells are starved they undergo meiosis leading to the production of haploid spores. Spores give rise to haploid cells of two different kinds of mating types **a** and α. These haploid cells can divide by budding, but when mixed they are able to mate and form diploid cells. Many changes are involved in the switch from haploid cells to diploid cells. The haploid cells produce a mating type specific pheromone, which diffuses into the medium and attracts mating partners of the opposite mating type. Diploid cells neither produce pheromone nor are they responsive to it. A G-protein coupled receptor system is involved in pheromone sensing in haploid cells and is one of the best studied signal transduction pathways in biology. The pheromone response includes increased transcription of genes whose products facilitate mating, arrest of the mitotic cell division cycle, changes in the cell surface and nucleus for fusion with the cognate organelles of the mating partner and alterations of the cell polarity and morphology (Sprague and Thorner, 1992).

We filtered for a *core set* of 20 genes known to be involved in pheromone response and their next neighbours in the disruption network. The subnetwork contains 63 genes and 115 arcs for $\gamma = 4.0$ (see Figure 4), 36 genes are adjacent to more than one arc. Of these 36 genes 18 genes belong to the core set and further 8 genes (STE4, STE5, STE7, STE11, STE12, STE18, STE24, KSS1) are annotated as involved in the mating response. Some of the remaining 10 genes are likely to encode proteins which are involved in processes that are related to the pheromone response. One example is HOG1, which encodes a MAP kinase like STE11, however this protein is involved in the high-osmolarity signal transduction pathway, which is similar to the pheromone response pathway. Another example is SHE4, which encodes a protein required for the duplication of the spindle pole body. Finding these groups could possibly be done by appropriate clustering. However, our method allows also to find genes with high outdegrees (SWI4, CYC8, TUP1, RPD3) in the subnetwork, which encode gene products known to be members of general regulatory complexes which influence many genes. The remaining four genes are RAD6, CRS4, YER044C and YER050C.

The same filtering at $\gamma = 2.0$ gives us a subnetwork with 240 genes and 470 arcs (see Figure 4), 75 genes have a degree higher than 1, including the full core set and the remaining genes mentioned in the previous paragraph. Additionally we find more genes encoding proteins which are likely to be involved in pheromone response related processes. For instance, BNI1 is required for the bipolar budding pattern, RTT104 is a helicase important for replication of ribosomal DNA, RAS2 is a GTP binding protein involved in regulation of the cAMP pathway, KIN3 is a serine/threonine kinase, FKS1 and GAS1 are involved in the cell wall synthesis.

The transcription factor Ste12 is the final protein in the pheromone signal transduction cascade. A consensus sequence for the DNA binding site of Ste12 is known ([A]TGAAACAA), and many genes involved in pheromone response have more than one Ste12-binding site in their promoter (Sprague and Thorner, 1992). Of 19 genes known to be induced by pheromone 11 are found to be connected to STE12 in the network for $\gamma = 2.0$ (10 for $\gamma = 3.0$), while only 2 out of 13 genes known not to respond to the pheromone signal, are connected to STE12 in the network with $\gamma = 2.0$ (none for $\gamma = 3.0$). We used PATMATCH[§] to locate the consensus sequence for the binding sites in the upstream regions of the genes (Vilo and Kivinen, 2001). The connectivity to STE12 in our networks corresponds to the absence or presence of the consensus binding site in the promoter regions of the corresponding genes (see supplementary data).

We tested several core sets containing genes involved in mating response with similar results. Up to 77% (depending on γ threshold) of the genes adjacent to the core set are annotated as involved in mating response as well.

We also tested core sets containing genes involved in other processes than mating response. These sets, e.g. small molecule transport, signal transduction and cell wall synthesis showed a similar behaviour, however fewer genes adjacent to the core set were involved in the same cellular process (up to 27% for small molecule transport, 20% for signal transduction, 21% for cell wall synthesis). For some core sets we found very few of the adjacent genes belonging to the same annotation group, e.g. energy generation with only up to 5% genes of the adjacent genes in the same group.

DISCUSSION

We studied gene disruption networks for yeast, inferred from gene expression data fully automatically without any human intervention. Yet, the overlap between these networks and the reference network constructed from the YPD database is always considerably higher than expected by chance. This encouraged us to study the properties of the disruption networks further.

We have studied structural features of these networks that are robust for a range of cutoff thresholds. We notice that the distributions of the numbers of incoming and outgoing arcs are rather uneven, with few genes having a large number of incoming or outgoing arcs, while most genes had very few of each. The genes with the highest number of outgoing arcs can be regarded as being ‘important’ for cellular regulation, and it is encouraging that the annotation of such genes indicate their regulatory functions. The genes with the highest number of incoming arcs can be regarded as having complex regulation, and again, it is encouraging that these genes typically have a metabolic function according to their annotations. ’Small molecule transport’ was the only class of genes that appeared in both classes.

We found that regardless of the cutoff threshold the disruption networks have only one clearly dominant connected component with rather small components of one to three genes being separated for higher significance threshold. Generally the same property stands if we remove the genes with the largest number of indegree or outdegree (these genes could be potentially be the ones holding the network together). The only case where the network falls into a number of subnetworks of relatively equal size is for very strong perturbations, and the disconnected componenents found when enforcing the network to break up this way do not have obvious biological meaning.

The network topology is dominated by this large connected component, and has a distribution of arcs which roughly follows a power-law. It has been suggested that this network topology is generated by a system which is optimised to work under conditions where it has to be robust against perturbations, but where this tolerance has a cost (Carlson and Doyle, 1999). This is typical for biological networks, which have gone through natural selection to be tolerant to uncertain environments. A biological system cannot be protected against all possible threats, since the cost would be too high. Therefore a topology is favoured where a disruption of one component is most likely to affect few others, but where it is unlikely to disrupt a central component which may cause severe damage or death. For instance protein networks (Jeong *et al.*, 2001) exhibit this type of robustness. Our findings support this theory since the data used is specifically information about gene disruptions. Intimately linked with this is the still open question about modularity in gene networks. For instance, Featherstone *et al.* argue that a network with a scale-free structure will be dominated by a single large component. Our results support this theory, since we were not able to find any discrete modules in our networks. However, it may be

[§] For PATMATCH and other tools within Expression Profiler see http://ep.ebi.ac.uk.

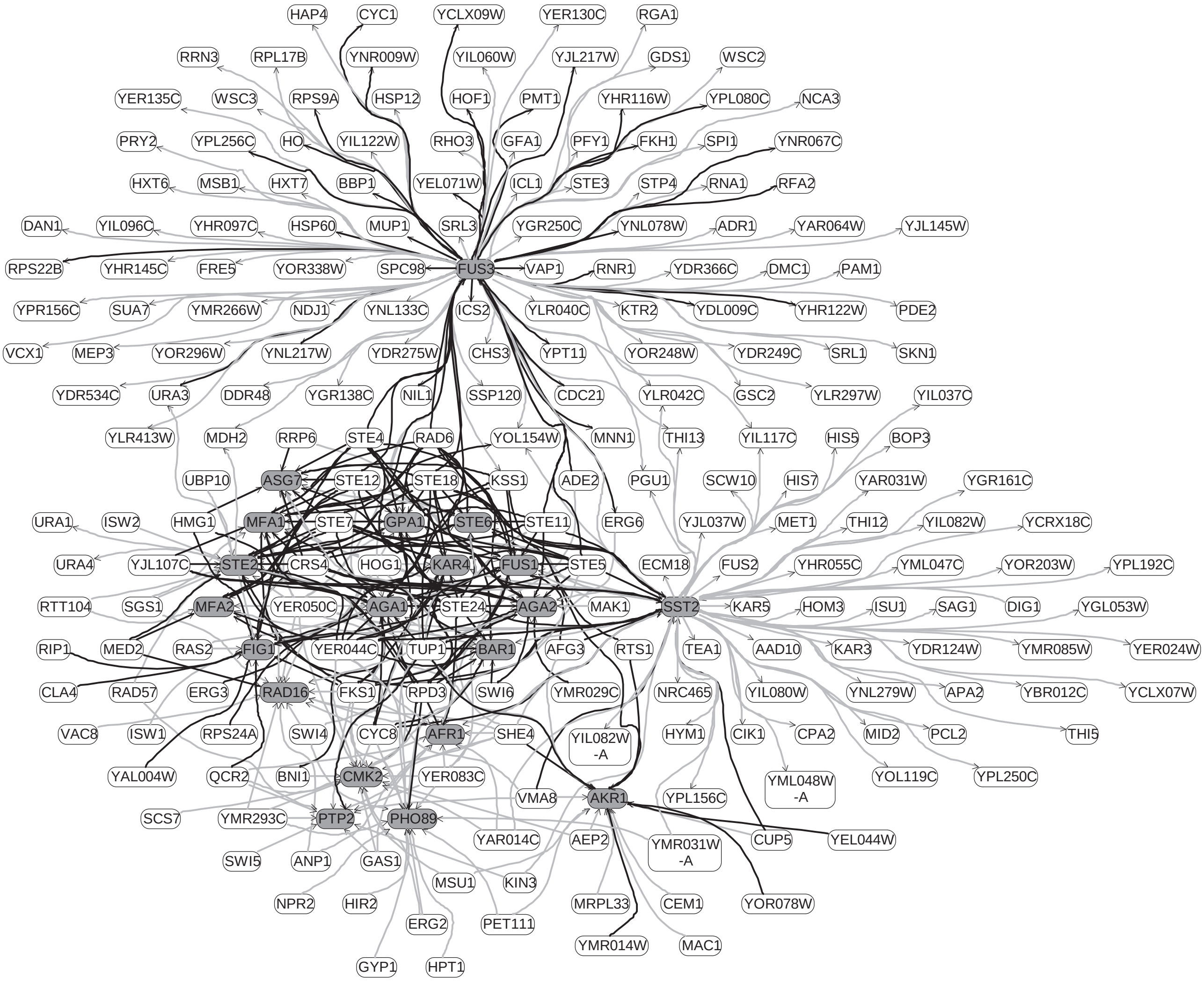

Fig. 5. This subnetwork is the result of filtering the full network at $\gamma = 2.0$ for the core set marked in grey and their next neighbours (grey arcs: downregulation, dark arcs: upregulation). See web supplement for a coloured figure (http://www.ebi.ac.uk/microarray/networks)

possible to find such modules by examining the networks in more detail, for instance by looking for the smallest cuts in the network graph which lead to disconnectivity and examine whether the resulting components have a biological meaning.

Looking for subnetworks consisting of genes involved in a particular cellular process allowed us to predict genes with similar functional roles. We need further studies to see how widely such predictions can be generalised.

Finally note that we can view the building of gene disruption networks as a method of structuring gene expression data which is an alternative to other known methods such as hierarchical clustering. Such networks allow us to explore different aspects of gene expression data.

ACKNOWLEDGEMENTS

We thank Frank Holstege, Patrick Kemmeren, Sabine Dietmann, Laurence Ettwiller and Michael Lappe for valuable discussions and Helen Parkinson for proofreading of the manuscript. JR acknowledges support from the AIM graduate school, Uppsala University, the Swedish Science Research Council (VR) and TTA Technotransfer AB.

REFERENCES

Albert,R. and Barabasi,A. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**, 47.

Barabasi,A. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509.

Carlson,J. and Doyle,J. (1999) Highly optimzed tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E*, **60**, 1412–1427.

Costanzo,M., Crawford,M., Hirschman,J., Kranz,J., Olsen,P., Robertson,L., Skrzypek,M., Braun,B., Hopkins,K., Kondu,P., Lengieza,C., Lew-Smith,J., Tillberg,M. and Garrels,J. (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.*, **29**, 75–79.

Featherstone,D.E.B.K. (2002) Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. *Bioessays*, **24**, 267–274.

Freivalds,K., Dogrusoz,U. and Kikusts,P. (2001) Disconnected graph layout and the polyomino packing approach. *Proc. of Graph Drawing*, in print

Freivalds,K. and Kikusts,P. (2001) Optimum layout adjustment supporting ordering constraints in graph-like diagram drawing. *Proc. Latvian Acad. Sci.*, **55**, 43–51.

Friedman,N., Linial,M., Nachman,I. and Pe’er,D. (2000) Using bayesian networks to analyze expression data. *RECOMB 2000*.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D., Kidd,M.J., King,A.M., Meyer,M.R., Slade,D., Lum,P.Y., Stepaniants,S.B., Shoemaker,D.D., Gachotte,D., Chakraburtty,K., Simon,J., Bard,M. and H.,F.S. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Jenssen,T.L., greid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, **28**, 21–28.

Jeong,H., Mason,S.P., Barabasi,A.L. and Oltavai,Z.N. (2001) Lethality and centrality inprotein networks. *Nature*, **411**, 41.

Jeong,H., Tombor,B., Albert,R., Oltavai,Z.N. and Barabasi,A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Murphy,K. and Mian,S. (1999) *Modelling Gene Expression Data Using Dynamic Bayesian Networks*, Technical Report, U.C. Berkeley, Department of Computer Science.

Pe’er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics* suppl 1, *ISMB 2001*, **17**, 215–224.

Sprague,G.J. and Thorner,J. (1992) Pheromone response and signal transduction during mating process of *Saccharomyces cerevisiae*. *The Molecular and Cellular Biology of the yeast Saccharomyces*, Volume 2 of *Monograph Series-21*, Jones,E., Pringle,J. and Broach,J. (eds), Cold Spring Harbour Laboratory Press, pp. 657–744.

Stapley,B. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualisation from co-occurrences of gene names in medline abstracts. *Pac. Symp. Biocomput.*, **5**, 526–537.

Vilo,J. and Kivinen,K. (2001) Regulatory sequence analysis: application to the interpretation of gene expression. *Eur. Neuropsychopharmacol.*, **11**, 399–411.

Wagner,A. (2002) Estimating coarse gene network structure from large-scale gene perturbation data. *Genome Res.*, **12**, 309–315.

Wolf,Y., Karev,G. and Koonin,E. (2002) Scale-free networks in biology: new insights into the fundamentals of evolution? *Bioessays*, **24**, 105–109.