Clustering and Cross-talk in a Yeast Functional Interaction Network

Jennifer Hallinan Anil Wipat
CISBAN & School of Computing Science,
University of Newcastle
Newcastle upon Tyne NE1 7RU

Abstract: **Many different clustering algorithms have been applied to biological networks, with varying degrees of success. The output of a clustering algorithm may be hard to interpret in biological terms because such networks are often large and highly interconnected, with structural and functional modules overlapping to varying degrees. In this paper we describe an evolutionary network clustering algorithm specifically designed for the analysis of large, complex biological networks. It identifies variably sized, overlapping clusters of nodes. The identification of points of overlap between clusters facilitates the analysis of the biological nature of crosstalk between functional units in the network. We apply two variants of the algorithm (one using probabilistic weights on edges and one ignoring them) to a recently published network of functional gene interactions in the yeast *Saccharomyces cerevisiae* and assess the biological validity of the resulting clusters in terms of ontological similarity.**

I. Introduction

Over the past decade developments in high-throughput biological assays, such as microarray, transcriptionally active polymerase chain reaction (TAP), proteomics and yeast two-hybrid protein interaction screens, have led to the generation of large amounts of data about biological interactions. The existence of these data sets means that large-scale biological interaction networks can be reconstructed and individual genes and gene products examined in the context of their genetic background. Analyses of various types of networks have produced unique insights into the behaviour of biological systems. For example, protein-protein interaction networks have been used to infer the protein function (e.g. Date & Stoeckert, 2006; Gavin *et al*., 2006) and transcriptional regulation (Drawid, Jansen & Gerstein, 2000), while metabolic networks have been used to investigate evolutionary relationships between widely divergent organisms in a manner not possible using single gene sequences (Jeong, Tombor, Albert, Oltvai & Barabasi, 2000).

Networks such as these are physical interaction networks, in which edges represent direct interactions such as protein-protein or protein-DNA binding. Although physical interaction networks are interesting in themselves, as discussed above, functional interactions within the cell are of even greater importance (Barabasi & Oltvai, 2004). Hartwell, Hopfield, Leibler & Murray (1999) argued that functional modularity is a critical level of biological organization. They define a functional module as "a discrete entity whose function is separable from those of other modules" (p.C48). Examples of functional modules include ribosomes, which are spatially isolated from other modules, or signal transduction systems, which are isolated by chemical specificity. Functional modules are frequently made up of heterogeneous agents interacting in a variety of ways, and hence will not be completely represented in a physical interaction network.

Network clustering has been the subject of extensive research. A network clustering algorithm aims to assign individual nodes to modules, the members of which are more tightly connected to each other than to the network in general. Information about cluster membership can be used to identify largely discrete structural or functional modules within a network, and provides clues to the possible function or location of unknown proteins, whilst the way in which modules are linked may reflect the overall functional organization of the network.

A wide range of biological interaction networks in numerous different organisms have been shown to be modular (Rives & Galitski, 2003; Thieffry & Romero, 1999; Hartwell, Hopfield, Leibler & Murray, 1999; Schuster, Pfeiffer, Moldenhauer, Koch & Dandekar, 2002; Han et al., 2004), and some are hierarchically modular, with small modules forming components of larger modules, which in turn are assembled into still larger modules (Holme, Huss & Jeong, 2003; Hallinan, 2004).

Functional modules are not necessarily isolated (Hartwell *et al.,* 1999); a given component may belong to different modules at different times, and the function of a module can be affected by signals from other modules. Such cross-talk between functional modules has been shown to be essential to the behaviour of a variety of different biological systems (e.g. Amin, 2004; Natarajan, Lin, Hsueh, Sternweis & Ranganathan, 2006). As with the identification of modules, the identification of the linkages constituting channels of cross-talk in a network is not straightforward. If there are many edges between two clusters, they merge into a single cluster,

but there is no obvious way of determining the cutoff above which this merger should happen.

Literally hundreds of clustering algorithms have been described (for an overview, see Hartigan, 1975), most of which can be modified to operate upon networks if a node distance metric can be specified. However, there are several drawbacks common to generic unsupervised clustering algorithms, particularly when applied to large, complex networks. Many algorithms, such as *k*-means and SOM, need to know the number of clusters in advance, and will partition data into the specified number of clusters whether or not that partitioning reflects real clustering in the network.

Hierarchical clustering algorithms are widely used because they are fast, provide a useful overview of the cluster structure of the network, and reflect the generally hierarchically modular nature of biological networks, but for practical use a decision must be made as to where in the cluster tree to threshold. This is not a straightforward decision, and often requires the use of further information about the network, which may not be available for all nodes of a large biological network.

Many clustering algorithms cannot use the information inherent in weightings on the edges in the graphs. While some biological networks, such as protein-protein interaction networks, are inherently unweighted, the edges in many networks represent interactions with which a metric can be associated. In metabolic networks, for example, kinetic parameters can be encoded as weights on edges between biochemical species. Clustering algorithms which do not incorporate weightings discard potentially valuable information about the network structure and function.

Most algorithms, whether or not they use a predetermined number of clusters, cluster all of the data provided. In many problem domains this is not an issue, but biological interaction networks inherently consist of structural and functional modules of varying sizes linked by nodes or short chains of nodes which lie, conceptually and topologically, outside the system of modules. Even more importantly, biological modules are essentially fuzzy, in that a single node may belong to more than one module, and modules may overlap to a greater or lesser extent in different parts of the same network. Biological networks also have a temporal element, with different modules likely to be active at different times and in response to different external stimuli.

In order to usefully cluster a large biological network, then, an algorithm should have the following characteristics:

- Ability to identify overlapping clusters of varying sizes;
- Requires no foreknowledge of number of clusters to be found;
- Does not necessarily assign all nodes to clusters;
- Requires no information about the network except topological structure;
- Can utilize weights on edges if they are present.

In this paper we describe the application of an evolutionary algorithm designed with these five criteria in mind, to what is probably the most complete functional interaction network published to date (Lee, Date, Adai & Marcotte, 2004). This network is of particular interest because it was constructed by integrating data from a number of sources, and concerns the model organism about which probably the most complete data sets exist, the yeast *Saccharomyces cerevisiae*.

II. Methods

A. The Yeast Functional Interaction Network

Lee *et al.* (2004) combined data on interactions between all yeast ORFs from 11 different sources, including protein-protein binding, regulatory, genetic and metabolic interactions. They used a Bayesian statistics approach to estimate the confidence level of each interaction, taking into account that for most interactions only some of the data sources were available. The Bayesian approach makes it possible to incorporate estimations of the reliability of the different data sets in a consistent manner. The result is a large network in which edges represent functional, and not necessarily physical interactions, and are weighted with the probability of that interaction. We used the 30,000 highest-ranking interactions in the Lee data set to generate a network.

B. Clustering Algorithm

The chromosome used in the evolutionary clustering algorithm is simply a string of integers, each representing an (arbitrarily numbered) node in a potential cluster. Both crossover and mutation operators were used. Populations size was 100 for each run.

The fitness of each individual reflects the relative number of edges between nodes in the cluster (k_i), and between nodes in the cluster and nodes outside the cluster (k_o). The fitness function was calculated using the cluster coherence measure, χ, described in Hallinan (2004):

$$\chi = \left(\frac{k_i}{n(n-1)} \right) - \frac{1}{n} \sum_{j=1}^{n} \left(\frac{k_{ji}}{k_{jo} + k_{ji}} \right) \qquad (1)$$

where k_i is the total number of edges between nodes in the module, n is the number of nodes in the network, k_{ji} is the number of edges between node j and other nodes within the module, and k_{jo} is the number of edges between node j and other nodes outside the module.

In order to allow the evolution of clusters of varying sizes, the chromosome length was allowed to vary. This was achieved by choosing independent breakage points for parents during crossover. Longer chromosomes potentially have higher fitness than shorter ones, simply because the number of potential edges within the cluster is larger. To counteract this tendency, the fitness of each chromosome was scaled by dividing it by the chromosome length.

The mutation operator involves the generation of a random number for each position in the chromosome. If the random number is less than a specified mutation rate, *mrate*, the

number in that position is replaced by an integer drawn with uniform probability from the range (1,*n*).

Since the functional network described by Lee et al. (2004) has confidence values associated with the edges, it was possible to implement a version of the genetic algorithm which took account of these edge weights. This was achieved simply by multiplying the number of edges k_i and k_o by the weight associated with each edge.

$$\chi = \left(\frac{2k_i w_i}{n(n-1)} \right) - \frac{1}{n} \sum_{j=1}^{n} \left(\frac{k_{ji} w_{ji}}{k_{jo} w_{jo} + k_{ji} w_{ji}} \right) \qquad (2)$$

As before, the fitness was scaled by chromosome length.

C. Cluster Validation

Although cluster analysis can be extremely useful in the analysis of large, complex data sets, validation of the resulting clusters is essential, in light of the known limitations of the approach (Handl, Knowles & Kell, 2005). In our algorithm cluster membership is assigned to nodes solely on the basis of network topology (and edge weighting, when appropriate). To validate the clusters we used the Gene Ontology (GO)[1] annotations for biological process, cellular component and molecular function.

III. Results

A. Whole Network Statistics

The statistics of the network, measured using the Pajek program (Batagelj & Mrvar, 1998) are shown in Table 1.

Table 1. Statistics of the network

Parameter	Value
Nodes	4,607
Edges	31,999
Average connectivity	6.95
Components	67
Maximum connectivity	183
Largest partition	4438 (96.3%)
Slope on line	-1.39
Diameter	16
Cluster coefficient	0.314

The structure of the complete network is shown in Figure 1. In this figure nodes represent individual genes, and edges are interactions between genes.

[1] http://www.geneontology.org/

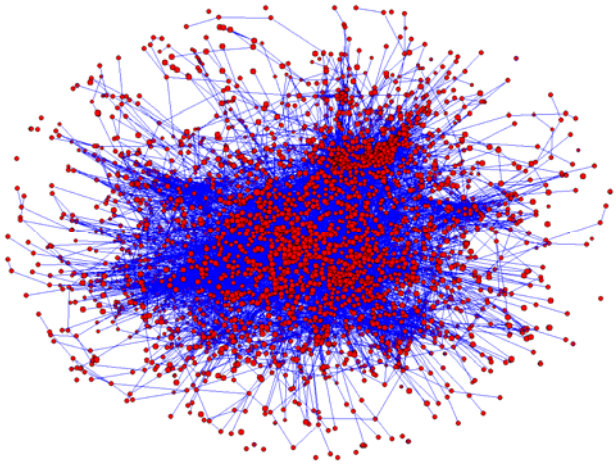

Figure 1. The probabilistic functional network generated from the data of Lee *et al.* (2004).

Although there are 67 connected components in the network, the majority (96%) of nodes are a single, giant connected component.

The connectivity distribution of the network is shown in a log-log plot in Figure 2. It is clear from Figure 2 that although there is an approximately linear portion to the log-log plot, the distribution is not convincingly scale-free, as many biological interaction networks have been demonstrated to be.

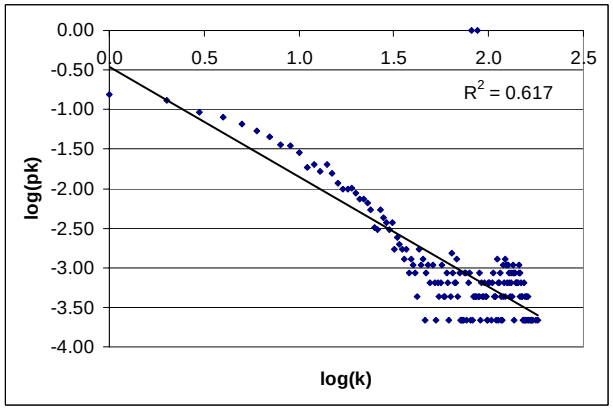

Figure 2. Log-log plot of the degree distribution. R^2 is 0.617

B. Algorithm Using Unweighted Edges

The version of the evolutionary clustering algorithm which did not incorporate edge weights was run 200 times, and the fittest individual from each run was recorded as a potential module. Fitness values ranged from 0.9268 to 0.5014. Out of the 4,407 nodes in the network, 490 were identified as participating in modules. There was considerable overlap between the modules detected with two nodes being present in 10 of the 200 modules.

The clusters found by the algorithm are shown in Figure 3.

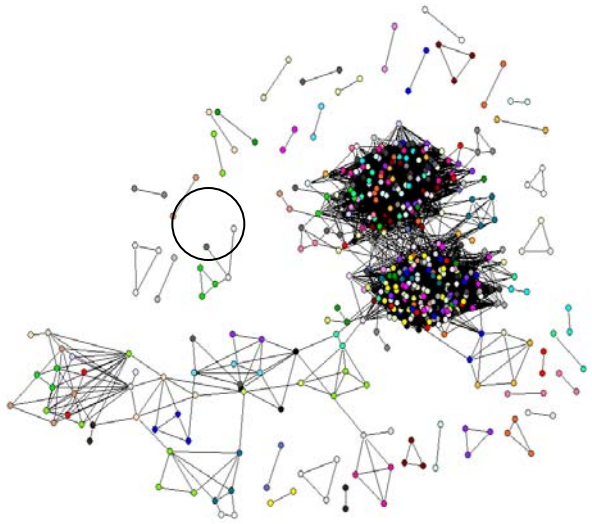

Figure 3. The subnetwork defined by modules detected by the unweighted algorithm.

Since the primary motivation for the generation of this clustering was comparison with the results of the algorithm which took account of the weights of the edges in the original network, the network shown in Figure 3. was not subjected to intensive analysis. The two large clusters of clusters correspond very closely with the largest two superclusters produced by the weighted algorithm and described in detail in Section C, below.

Two large “superclusters” are apparent, together with a number of other more-or-less overlapping superclusters, plus several distinct clusters. All of the two- and three-node clusters in Figure 3 are single clusters, except for the one ringed.

In order to identify nodes linking the superclusters we used Freeman's measure of betweenness centrality (Freeman, 1977). This metric reflects the extent to which nodes lie on the geodesics of the network (the shortest paths between a pair of nodes). Nodes of high betweenness are good candidates as mediators of cross-talk between modules, since by definition they act as bridges between large numbers of paths through the network.

The betweenness distribution of the nodes in the network is shown in Figure 4.

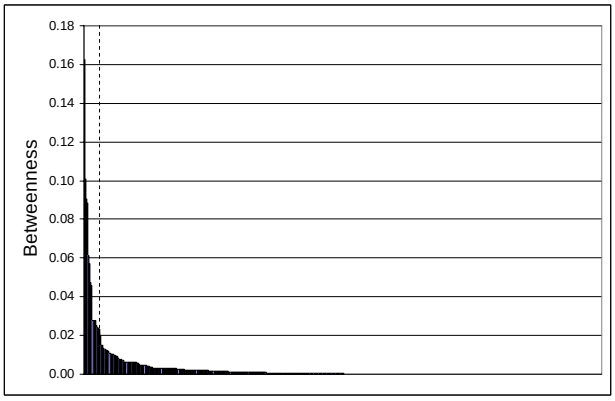

Figure 4. Betweenness distribution of the unweighted clusters. The dashed line indicates the cutoff (0.02) above which nodes are considered to have high betweenness.

Inspection of the distribution reveals a sharp drop at a value of 0.02 (dashed line). Nodes with betweenness above this value were designated "high betweenness" (Figure 5).

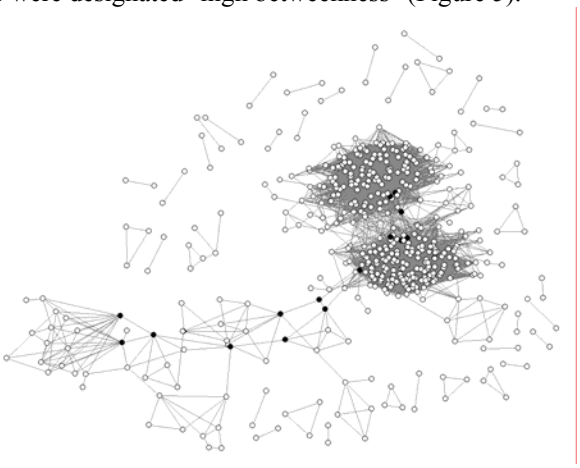

Figure 5. Nodes of high betweenness are shaded black

C. Algorithm Using Weighted Edges

Visual inspection of the results of the algorithm incorporating weights on the edges between nodes in graph indicated that many of the runs had failed to find connected clusters (Figure 6). There appears to be a clear demarcation between runs which achieved a fitness of 0.4 and over, and those which did not. “Clusters” with fitness less than 0.4 tended to consist of scattered individual nodes, not linked to any other node. Since these clusters are clearly spurious, only clusters with fitness greater than or equal to 0.4 were used in the final analysis.

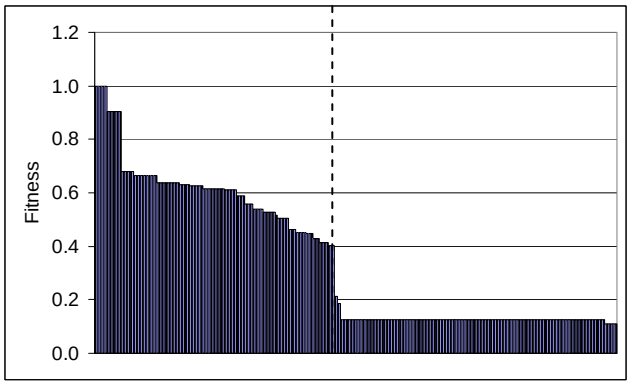

Figure 6. Fitness in the results of the weighted algorithm. The dashed line indicates the cutoff (0.4) above which runs were considered to have been successful.

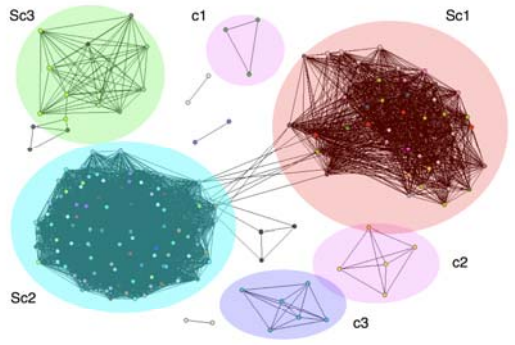

Figure 7. Clusters detected by the algorithm using weighted edges.

Application of the algorithm to the weighted network, identifies two major aggregations of clusters, in addition to a number of smaller aggregations and single clusters comprising pairs or triples of nodes (Fig. 7). The two major aggregations, termed supercluster 1 (Sc1) and supercluster 2 (Sc2), comprise 59 genes arranged in 19 clusters, and 98 genes in 30 clusters, respectively. Sc1 and Sc2 are connected by 10 edges. A third aggregation, termed supercluster3 (Sc3) comprises 18 genes in 7 clusters. Six other single clusters, designated c1 to c6, of two or more proteins are also defined. Sc3 and the small clusters are isolated from Sc1 and Sc2. Sc1 and Sc2 show tight functional coherence and contain genes encoding proteins responsible for the synthesis of ribosomes (ribosomal biogenesis) and genes encoding the structural proteins themselves, respectively.

Ribosomes are cellular structures composed of two subunits, each a complex of protein and RNA. They play a central role in the process of translation. A typical cell, under optimal growth conditions, contains enormous numbers of ribosomes. Whilst ribosomal biogenesis and assembly have been extensively studied, they are extremely complex and coordinated processes. Our knowledge of these processes is still incomplete and the role of many of the proteins involved remains to be determined.

The synthesis of ribosomes in eukaryotes involves the processing of a large pre-ribosomal RNA molecule to form the structural RNA molecules onto which a large number of ribosomal proteins are subsequently assembled. Recent studies in *S. cerevisae* have revealed that over 200 proteins are involved in synthesis alone (distinct from structural proteins), giving rise to a number of intermediate complexes (Dlakić, 2005). For an extensive review of ribosomal biogenesis in *S. cerevisae*, see Granneman and Baserga (2004).

Composition of Sc1

Sc1 mainly comprises genes involved in the biogenesis and synthesis of ribosomes. As expected, the proteins encoded by these genes are mostly located in the nucleus, within a defined structure called the nucleolus (Table 2).

TABLE 2. GO BIOLOGICAL PROCESS SUMMARY FOR SC1

Process	**Count**
rRNA processing	8
Biological process unknown	8
35s primary transcript processing	7
Ribosomal large subunit biogenesis	6
Ribosomal large subunit assembly and maintenance	5
Processing of 20S pre-rRNA	5
tRNA methylation	3
Telomere maintenance	2
tRNA modification	2
Vesicle fusion	1
Ubiquitin-dependent protein catabolism	1
Transcription from RNA polymerase II promoter	1
transcription from RNA polymerase I promoter	1
Ribosome assembly	1
Response to osmotic stress	1
regulation of transcription, mating-type specific	1
protein import into mitochondrial matrix*	1
peroxisome organization and biogenesis*	1
mRNA export from nucleus	1
meiosis	1
chromatin silencing at telomere*	1
actin cytoskeleton organization and biogenesis*	1
Total	59

In *S. cerevisiae*, pre-RNA molecules are transcribed in the nucleus by distinct RNA polymerases to generate the large polycistronic 35S pre-rRNA and the shorter 5S rRNA molecules that are ultimately processed by a number of complex cleavage steps into the shorter mature ribosomal

RNA (rRNA) molecules (5.8S, 25S/28S and 5S rRNA for the 60S subunit and 18S for the 40S subunit) (Fromont-Racine *et al*., 2003). The process of synthesis involves the action of a number of enzymes including endo and exonucleases which cleave RNA, and enzymes which extensively modify the RNA molecules themselves. The modification of rRNA is mediated by small nucleolar RNA fragments (snoRNA's), which are thought to direct the action of RNA modifying enzymes. snoRNA's exist in form complexed with a number of proteins, some of which possess the necessary enzymatic activity for RNA modification. Examination of the clusters that make up Sc1 reveals that they are mostly clusters of genes encoding a range of RNA processing and modifying enzymes together with genes encoding the proteins that complex with snoRNA's (data not shown).

Composition of Sc2

The clusters comprising Sc2 show remarkable uniformity in terms of their predicted function (Table 3).

TABLE 3. GO BIOLOGICAL PROCESS SUMMARY FOR SC2

Process	Count
protein biosynthesis*	86
tRNA export from nucleus*	1
telomere maintenance*	1
phosphate transport	1
mRNA export from nucleus	1
methionyl-tRNA aminoacylation	1
meiosis	1
low-affinity zinc ion transport	1
chromatin silencing at telomere*	1
chromatin modification	1
biological process unknown	1
aldehyde metabolism	1
35S primary transcript processing*	1
Total	**98**

In eukaryotes, such as yeast, the 40S subunit contains a single 18S rRNA and 32 different ribosomal proteins, while the 60S subunit contains the 5S, 5.8S, and 25S rRNAs and 48 different ribosomal proteins. Most of the genes in the clusters Sc2 encode these ribosomal structural proteins.

Composition of Sc3

Sc3 contains clusters of genes, about two thirds of which are of unknown function (Table 4).

TABLE 4. GO BIOLOGICAL PROCESS SUMMARY FOR SC3

Process	Count
biological process unknown	8
vacuolar protein catabolism	1
triacylglycerol biosynthesis*	1
telomere maintenance*	1
sulfur metabolism	1
response to oxidative stress	1
response to metal ion	1
protein targeting to vacuole	1
methionine metabolism*	1
fatty acid oxidation	1
ER to Golgi vesicle-mediated transport	1
Total	**18**

The member genes that have been characterised are functionally diverse, ranging from a putative transcription factor, through enzymes involved in lipid metabolism. Preliminary analysis of Sc3 or its subclusters does not reveal any clues about the potential role of this supercluster.

Betweenness and connections between the sc1 and sc2.

Inspection of the betweenness distribution of the weighted network led to the selection of 0.02 as the cutoff above which nodes were designated "high betweenness" (data not shown). Nine nodes were selected (Figure 8).

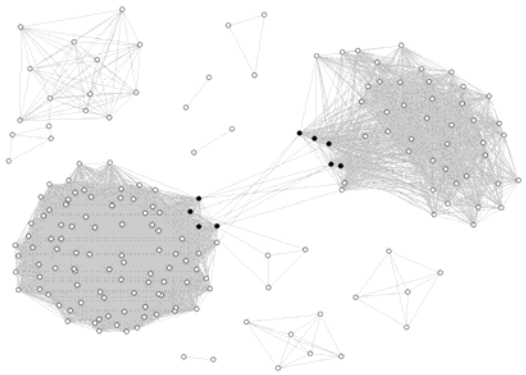

Figure 8. Nodes of high betweenness in the weighted network (shaded in black).

The selected nodes and the proteins for which they code are shown in Table 5.

TABLE 5. NODES OF HIGH BETWEENNESS IN THE WEIGHTED NETWORK

Gene	Protein
YOL121C	40S small subunit ribosomal protein S19.e
YGR214W	40S ribosomal protein p40 homolog A
YIL133C	60S large subunit ribosomal protein
YJL189W	60S large subunit ribosomal protein L39.e
YGR128C	hypothetical protein
YLR196W	similarity to human IEF SSP 9502 protein
YNL002C	strong similarity to mammalian ribosomal L7 proteins
YGL078C	putative RNA helicase required for pre-rRNA processing
YKL009W	mRNA turnover 4

IV. Discussion

The full network is, as expected, more highly connected than most physical interaction networks, with an average connectivity of nearly 7, whereas most physical interaction networks have been shown to have average connectivity of around 2 – 3. This increase reflects the multiple nature of the data from which the network is constructed. The network is also less of a good fit to a scale free distribution than many physical interaction networks described in the literature, probably for the same reason. These characteristics suggest that modules in the network are likely to be less clearly distinguished from their background than those in, for example, a protein-protein interaction network.

As expected, there are marked differences between the subnetworks produced by the weighted and unweighted algorithms. The unweighted algorithm, which treats all edges as equally important no matter what the evidence for their existence, produces a network in which there are two major clusters of genes and many small "clusters" consisting of two or three nodes: a total of 43 connected components. The weighted algorithm, in contrast, produces a subnetwork with only 20 connected components, including the two large clusters identified by the unweighted algorithm.

The general structure of the networks produced by the two versions of the algorithm were similar. Both identified two large "superclusters" of overlapping clusters, one of which contained ribosomal structural genes, and one of which comprised rRNA processing genes. These clusters were the same as the largest two clusters identified by the hierarchical clustering algorithm used by Lee *et al.* (2004). Since our clustering is incomplete in that it covers only a small proportion of the nodes in the complete network, it is not surprising that the two largest clusters are over-represented. The weighted algorithm identified a smaller and more functionally coherent set of clusters and superclusters, indicating that the use of edge weights by a clustering algorithm, where available, is likely to be valuable in producing biologically relevant results.

Lee *et al*. (2004) describe the network as " a highly modular gene network with well-defined subnetworks." (p. 1556). They used a hierarchical clustering algorithm to identify clusters in the network, and then delineated individual clusters on the basis of a coherence measure based upon their functional annotation. In contrast, our aim was to use only topological structure for the clustering, reserving GO annotations for cluster validation. They identified clusters of genes involved in energy metabolism, DNA damage response and repair, mitochondrial ribosome, ribosome, ribosome biogenesis, cellular transport, mRNA splicing and chromatin modelling, of which the two largest are ribosome and ribosome biogenesis genes. These were also the two largest clusters identified by our algorithm.

The domination of the clustering by two sets of genes which are well known to be closely functionally interrelated, although understandable, suggests that the approach to network construction used here may have some drawbacks.

Edges were selected for inclusion in this network on the basis of their computed probability. Following Lee *et al.* the top 30,000 most highly weighted interactions were used. This approach risks "swamping" the clustering algorithm with tightly linked, but not particularly interesting, clusters. To detect and examine less easily predicted clusters it may require the discard of some of the most heavily weighted edges, on the grounds that they are unlikely to contribute novel information. Our laboratory is currently exploring the development of a metric for this purpose, along the lines of those used in linguistic analysis, where common words are generally discarded from the corpus prior to analysis.

The topological rationale for the identified nodes of high betweenness is clear; all high betweenness nodes occur in one of the two largest superclusters, and have large numbers of edges with other nodes in the same supercluster and only one or two edges with links in the other major supercluster. However, investigation of their biological function, as evidenced by GO annotation, reveals no clear biological reason why those proteins should be involved in crosstalk between the two superclusters. There are two possible explanations for this observation. The first possibility is that the GO annotation is simply incomplete, and the proteins in question have biological functionality which has not yet been identified experimentally. This suggestion is supported by the fact that five of the nine nodes of high betweenness are annotated as "putative" or "homolog".

The other potential explanation concerns the evidence upon which the existence and weight of the link is inferred. For several of the edges associated with nodes of high betweenness the only evidence is co-expression, as detected by DNA microarrays. Although co-expression may indeed reflect a direct functional relationship, such as co-regulation, this is not always the case; Allocco, Kohane & Butte (2004) found that genes which were co-expressed had a greater than 50% chance of sharing a common transcription factor only when the correlation between expression profiles was greater than 0.84. Edges which exist in the network by virtue of co-expression evidence might represent relationships which are simply not apparent from the GO annotation of their associated nodes.

Both of these explanations probably apply to our results. If correct, this conclusion necessitates re-assessment of the underlying assumptions of the topological analysis of this particular network. The primary assumption is that analysis of a network of functional, as opposed to strictly physical, interactions will provide higher-level information about the way in which the cellular components are organized. Like us, Lee *et al.* (2004) used a clustering algorithm primarily based on topology to cluster their network into linked clusters of genes involved in related functions. They did not attempt to investigate the biological meaning of the links in the same way in which they explored the biological significance of the gene annotations.

The Bayesian algorithm used by Lee and colleagues to assign probabilistic weights to edges was designed to take into account the inherent reliability (incorporating factors such as noise and error rate) of the experimental data used. We contend that this approach, while a valuable advance on most previous work, which has made no attempt to integrate diverse data sources in a principled manner, is still insufficient to enable analysis of the nature of crosstalk between functional models. The existence of nodes and edges linking functional modules is easily confirmed, but a useful biological interpretation of this topological data will require still more sophisticated metrics for data integration.

The results discussed here provide a proof of principle for the value of the evolutionary clustering approach used, rather than a comprehensive analysis of the yeast functional interaction network. Only 200 runs of each version of the algorithm were performed, each of which returned a single cluster of, on average, eight genes. Each analysis as performed could therefore have examined no more than about 800 of the 4,760 genes in the network, and because of the desired overlap between clusters actually clustered far fewer. This partial analysis has, however, demonstrated that the evolutionary clustering algorithm is valuable because a) it reliably identifies overlapping clusters of nodes of varying sizes; and b) is produces more biologically valid clusters when it incorporates the probability values on the edges when computing the coherence of the clusters. Work is ongoing in our laboratory to fully characterize the yeast network generated by Lee *et al.* (2004), and to extend this approach to the functional networks of other organisms.

Acknowledgments

This research was funded by the Biotechnology and Biological Sciences Research Council through the Centre for the Integrative Systems Biology of Ageing and Nutrition (CISBAN) and the Newcastle University Systems Biology Resource Centre.

References

[1] Allocco, D. J., Kohane, I. S. & Butte, A. J (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* 5(18) http://www.biomedcentral.com/1471-2105/5/18 Downloaded 30/05/2006.

[2] Amin, A, (2004). Genetic cross-talk during head development in Drosophila. *Journal of Biomedicine and Biotechnology* 2004(1): 16 - 23.

[3] Barabasi, A.-L. & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5: 101 - 114.

[4] Batagelj, V. & Mrvar, A. (1998). Pajek - Program for Large Network Analysis. *Connections* 21: 47 - 57.

[5] Date, S. V. & Stoeckert, C. J. (2006). Computational modelling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. In Date, S. V. DOI:10.1101/gr.4573206. *Genome Research.*

[6] Dlakić, M. (2005). The ribosomal subunit assembly line. *Genome Biology* 6(10): 234 - 242.

[7] Drawid, A., Jansen, R. & Gerstein, M. (2000). Genome-wide analysis relating expression level with protein subcellular localization. *Trends in Genetics* 16(10): 426 - 430.

[8] Freeman, L.C. (1977). A set of measures of centrality based on betweenness. *Sociometry* 40(1): 35 - 41.

[9] Fromont-Racine, M., Senger, B., Saveanu, C. & Fasiolo, F. (2003). Ribosome assembly in eukaryotes. *Gene* 313: 17 - 42.

[10] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M. & Marzioch, M. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440: 631 - 636.

[11] Granneman, S. & Baserga, S. J. (2004). Ribosome biogenesis: Of knobs and RNA processing. *Experimental Cell Research* 296: 43 - 50.

[12] Hallinan, J. (2004). Gene duplication and hierarchical modularity in intracellular interaction networks. *Biosystems* 74(1 - 3): 51 - 62.

[13] Han, J-D. J., Bertin, N., Hao, T., Goldberg, D. T., Berriz, G. F. & Zhang, L. V. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430: 88 - 93.

[14] Handl, J., Knowles, J. & Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* 21(15): 3201 - 3212.

[15] Hartigan, J. A. (1975). *Cluster Analysis*. New York: Wiley.

[16] Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402(Supp): C47 - C52.

[17] Holme, P., Huss, M. & Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics* 19: 532 - 538.

[18] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabasi, A.-L. (2000). The large-scale organization of metabolic networks. *Nature* 407: 651 - 654.

[19] Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. (2004). A probabilistic functional network of yeast genes. *Science* 306(5701): 1555 - 1558.

[20] Natarajan, M., Lin, K.-M., Hsueh, R. C., Sternweis, P. C. & Ranganathan, R. (2006). A global analysis of cross-talk in a mammalian cellular signalling network. *Nature Cell Biology.* Advance Online Publication DOI: 10.1038/ncb1418.

[21] Rives, A. W. & Galitski, T. (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the USA* 100(3): 1128 - 1133.

[22] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. & Dandekar, T. (2002). Exploring the pathway structure of metabolism: decomposition into subnetworks and application to *Mycoplasma pneumoniae. Bioinformatics* 18: 351 - 351.

[23] Thieffry, D. & Romero, D. (1999). The modularity of biological regulatory networks. *Biosystems* 50: 49 - 59.