Clustering Incorporating Shortest Paths Identifies Relevant Modules in Functional Interaction Networks

Jennifer Hallinan, Matthew Pocock, Stephen Addinall, David A Lydall and Anil Wipat

Abstract—Many biological systems can be modeled as networks. Hence, network analysis is of increasing importance to systems biology. We describe an evolutionary algorithm for selecting clusters of nodes within a large network based upon network topology together with a measure of the relevance of nodes to a set of independently identified genes of interest. We apply the algorithm to a previously published integrated functional network of yeast genes, using a set of query genes derived from a whole genome screen of yeast strains with a mutation in a telomere uncapping gene. We find that the algorithm identifies biologically plausible clusters of genes which are related to the cell cycle, and which contain interactions not previously identified as potentially important. We conclude that the algorithm is valuable for the querying of complex networks, and the generation of biological hypotheses.

I. INTRODUCTION

Modern molecular biology often involves the application of high-throughput experimental procedures which generate large amounts of data. Technologies such as microarrays, yeast two-hybrid screens and tandem affinity purification provide insights into the biology of individual genes and proteins, and also into the web of interactions between them. Over the past decade or so, the study of biological interaction networks has become an important part of systems biology.

Networks may be built using data from a single type of experiment, such as protein-protein interactions, or data from multiple sources may be integrated into a single network. In a single-source network, the amount of information contained in an edge is limited and clearly interpretable. If there is an edge between two genes their products physically bind to each other under certain experimental conditions (which may not, of course hold in the living organism; the extent of error in these networks is very difficult to quantify [1]). The advent of integrated functional networks, however, raises the possibility of inferring more information from edges.

A. Integrated Functional Networks

An integrated functional network is constructed from data produced by multiple types of experiment, combined using a statistical approach to yield a network in which nodes represent genes and edges reflect any sort of interaction between genes or their products. The edges are weighted to reflect the likelihood of the existence of that edge, reflecting the different amounts of error in the data sets from which they are composed. In an integrated functional network, whilst the presence of an edge indicates the existence of some kind of functional relationship, the biological interpretation of that relationship is not necessarily clear. The different data sources comprising an integrated network provide different amounts and types of information, and combining them is not a conceptually straightforward process [2].

Integrated functional networks are increasingly being used as tools to analyze the function and relationships of subsets of genes, identified via high-throughput laboratory experiments [3]. Microarray experiments, for example, often yield a list of a small number of genes whose patterns of expression indicate that they may be important in the condition under investigation. These genes can then be located in a larger integrated interaction network and a number of questions asked, such as: do these genes, identified on the basis of phenotype, interact in any way? What other genes do they interact with? Are they part of a single functional or structural module or pathway?

The most widely used approaches to these problems to date have involved either examining the neighborhoods of genes—the subsets of genes to which those of interest are immediately connected—or applying a network clustering algorithm to identify tightly connected clusters of genes to which the genes of interest belong. In both cases the basic premise is that of "guilt by association": that the characteristics of genes or proteins can be inferred from the company they keep. Analyses such as these are global. The topology of the entire network is of interest, and summary statistics such as network diameter and cluster coefficient are calculated over all of the nodes. Sometimes, however, only a subset of the network is of interest.

B. Network Clustering and Modularity

Many networks appear to be organized into a number of modules. A module is generally defined as a subnetwork of a graph, the nodes of which have more connections to other nodes within the module than to external nodes [4]–[7].

The identification of modules within a network is an NP-complete problem [8]. It is, however, an important problem; Newman [9] describes it as "one of the outstanding issues in the study of networked systems". This is particularly true in the case of integrated functional networks. Although physical interaction networks are interesting in themselves, functional interactions within the cell are of even greater importance [10]. Hartwell, Hopfield, Leibler & Murray [11] 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. In bacteria, Spirin et al [12] have shown that evolutionary modules, rather than pathways, seem to make up the regulatory and functional units. Similarly, Campillos et al [13], identify evolutionarily cohesive modules enriched in genes for functions such as energy metabolism, across a number of prokaryotes. On a dissenting note, Wang & Zhang [14] argue that because there does not appear to be significant evolutionary conservation between modules in protein-protein interaction networks in yeast, fly and nematode, it is possible that the structural modules of the network originated as "an evolutionary by-product without biological significance". Since their investigation only included physical interaction data; it would be of considerable interest to repeat this analysis with an integrated functional network.

Functional modules are not necessarily completely isolated [15]; 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 [16][17].

A number of algorithms have been used for the identification of modules in networks. Some are based upon statistics and graph theory, while others are adapted from clustering algorithms developed for non-network data. It has recently been suggested that in addition to a modular organization, biological networks tend to have a hierarchical structure, in which nodes are organized into small modules which are, in turn, organized into larger modules, and so on [7]. Hierarchical modularity has, indeed, been identified in the metabolic networks of 43 different organisms [18].

Although several clustering approaches have been successfully applied to biological interaction networks, there are a number of issues specific to integrated functional networks which have not been fully addressed. 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 previous work we have developed a genetic algorithm-based approach to clustering integrated functional networks [19], which overcame some of the problems associated with the identification of functional modules. When applied to an integrated functional network for the yeast Saccharomyces cerevisiae, it identified individual clusters which tended to overlap into coherent superclusters. The cluster structure of the network was, however, dominated by large, well-characterized clusters of genes and proteins involved in the essential housekeeping tasks of ribosomal biogenesis and protein biosynthesis.

Many of the high-throughput, genome-wide experimental techniques developed over the last decade or so are not designed to investigate a specific hypothesis. Experiments such as DNA microarrays [20] or yeast two-hybrid screens [21] tend to be carried out in an exploratory manner, to investigate what happens within the cell under certain specific conditions. The result is often a list of "genes of interest", which are known to be involved in the phenomenon under investigation (e.g. genes which are upregulated in a specific type of cancer), but whose biological function and relationships may be obscure. These genes and those with which they interact closely are likely to be of more interest to the investigator than the rest of the genes in the organism. For the purposes of querying with a set of genes of interest, it is valuable to have some way of filtering out much of the network, allowing the researcher to focus upon the subset of genes and interactions most likely to be functionally related to the genes of interest.

If a network is being queried with a list of genes of interest, the topological closeness of other genes in the network to the genes of interest may be valuable information. We propose a measure of relevance, based
upon the topology, and particularly the weighting, of a probabilistic network. We refer to this measure as "topological relevance". It involves two assumptions: that nodes which are close to a node of interest are likely to be of interest; and that more distant nodes are also likely to be of interest, if the paths which connect them to a node of interest are sufficiently highly weighted.

C. Relevance

Edge weights are normally interpreted as reflecting the probability of existence of an edge. A less obvious weighting which could be applied to edges is that of relevance to the question being investigated. Since different data sources provide different types of information, it follows that for specific questions some data sources will be more relevant than others. For the analysis of a signal processing pathway, for example, protein-protein interaction data is likely to be more relevant than coexpression information, since genes in signaling pathways are often expressed at basal levels.

Relevance, however, is not an easy concept to define, let alone quantify. The Compact Oxford English Dictionary defines 'relevant' as meaning "closely connected or appropriate to the matter in hand". Under this definition, the relevance of a particular data set depends largely upon the question that is being asked. Relevance is also used in this sense by designers of algorithms for web search engines, who aim to return pages whose content is relevant to the query terms used. Such algorithms tend to be carefully guarded secrets. In an integrated network relevance should be able to be represented as a weighting, which can be incorporated with other weightings into the equation used to calculate the probability of an edge.

Several workers have tried to incorporate the concept of relevance into the analysis of integrated functional networks. One approach which has been tried is to use an expert to provide a subjective value of the relevance of different data types [22][23]. Although this is an obvious first approach to try, it is somewhat arbitrary. Further, these authors did not make any comparison of their algorithms on relevance-weighted versus non-relevance-weighted networks, and so the effect of the relevance weighting was not apparent.

A slightly different definition of relevance is taken by Antonov et al. [24]. These researchers assume that the relevance of a network can be measured as a bias on the weights of edges. Different data sets, they argue, will weight different types of edges highly, and measurements of this tendency can be used to give an estimate of the global relevance of a particular network. Their BioRel web site provides tools to compute the relevance of networks for a number of species. This is an interesting approach, and provides useful information about the characteristics of large biological networks. However, their use of the term 'relevance' as a synonym for 'bias' does not reflect our definition as given above.

All of the work described so far is based upon the assumption that the relevance of a data set to a particular question is reflected in the weight on its edges. Edge relevance is not, however, the only sort of relevance which is valuable to include in an integrated functional network. Nodes may also vary in their relevance to a particular investigation, either because of some characteristic of the gene which they represent, or by virtue of their position within the network.

In this paper we develop a genetic algorithm incorporating topological relevance as part of the fitness function, and use it to investigate an integrated functional network of yeast genes for clusters of nodes relevant to a set of genes identified as interesting in a separate genome-wide screen.

II. METHODS

A. Identifying Genes of Interest

Telomeres, nucleoprotein structures which protect the ends of chromosomes, appear to be important to the cell cycle and the ageing process in humans. With every cell division the telomeres become shorter, until eventually they are too short to bind the DNA replication machinery. The cell stops dividing, senesces and eventually dies. Although the overall picture of role of telomeres is clear, the pathways controlling telomere maintenance and elongation are still incompletely understood.

In an effort to systematically identify genes and proteins important in telomere maintenance, Downey et al. [25] carried out a genome-wide screen in Saccharomyces cerevisiae for suppressors of a mutated version of the gene cdc13, encoding a protein known to be important in telomere capping.

A library of strains of yeast, each carrying a temperature-sensitive mutant form of cdc13 together with one other knocked-out gene, was constructed, incubated at both permissive and non-permissive temperatures, and the growth of colonies assessed. Genes which "rescued" their carriers from the effects of the cdc13 mutation were identified on the basis of their increased growth compared with controls. The identification of genes potentially interacting with cdc13 was therefore made on the basis of a common phenotype (i.e. a common increase of growth).

There were 51 genes identified as "of interest" by the high throughput screen. Further laboratory analysis of this subset of genes identified a protein complex known as the KEOPS complex, which promotes both telomere uncapping and telomere elongation [25].

The degree distribution of this subset of genes is shown in Figure 1.
Figure 1. Degree distribution of the 51 genes identified as of interest in the high throughput screen.

The degree distribution of this set of genes is of interest because many biologically-based networks tend to have a scale-free degree distribution, with a large number of nodes of low degree, and a small number of highly-connected "hubs". The presence of hubs has been hypothesized to fundamentally affect the stability and information-transfer properties of network [26]. This particular set of genes does not conform to this pattern, having a higher than expected number of well-connected hubs.

Eighteen of the fifty-three genes form a single, highly-connected component in the graph (Fig. 2). The other genes tended to occur singly.

Fig. 2. Connectivity of the group of genes of interest. There is a single large connected component and a number of smaller ones.

Since it appears likely that genes in the largest connected component (LCC) may share a degree of biological function or process, this subset of genes was initially used to query the network.

B. The Yeast Integrated Functional Network

The baker's yeast, Saccharomyces cerevisiae, is undoubtedly the most intensively studied eukaryotic model organism, at the genome-wide level. Information about the relationships between genes and gene products has been generated by many workers, using a wide variety of high-throughput, genome-wide techniques, such as microarrays [27][28], yeast two-hybrid assays [29] and other protein interaction analyses [30], transcription factor binding [31][32] and many others. Whilst analysis of these networks in isolation has proven valuable for tasks such as assigning function to unknown proteins, there is growing interest in the topic of integrating interaction networks from different sources to produce a single network of functional interactions. Such a network was produced by Lee et al. for S. cerevisiae [33]. This network, comprised of data from 11 different sources, has nodes representing proteins and edges representing one or more interactions between a pair of proteins. The edges are weighted, using a Bayesian approach, to represent the probability that an edge exists in reality.

We sorted the interactions in the Lee network by probability of occurrence, and selected the top 32,000 interactions. The resulting network consists of 4,607 nodes and 32,000 edges, giving an average connectivity of 6.9. The connectivity distribution was, however, close to scale-free, a phenomenon which has been observed for many biological networks (Fig. 3).

C. Fuzzy Genetic Clustering Algorithm

The clustering algorithm described here is an extension of that described in [19]. It is a simple genetic algorithm, in which each individual consists of a string of integers, each representing a node in the network. The nodes comprising a single individual are interpreted as the members of a single cluster, and their fitness calculated as described below. Both random mutation (to any other valid node in the network, since there is no topological significance to the node numbering scheme) and two-point crossover are used.

The two-point crossover means that individual chromosomes can grow and shrink in length, encoding larger or smaller clusters. The probability that two individuals selected as parents would participate in crossing over was controlled by a parameter, the crossover rate. For each pair of parents a random number in the range (0,1) was generated, and crossover occurred if that number was smaller than the crossover rate. A new random number in the range (1, chromosome length) was then generated for each parent in order to determine the crossover point for that individual.

Tournament selection was used, with retention of the single fittest member of the current population.

The fitness function was based upon a measure of
modular coherence (eq. 1).

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

The coherence measure can take values in the range (-1,1), with 1 representing a fully connected, stand-alone module.

In the genetic algorithm described in [19] the edges, \( k \), were weighted by the probabilities of each edge existing. In the current algorithm, the edge weights were not included in the coherence function, but instead incorporated into a measure of topological relevance which was assigned to each node.

This was done by weighting each node in the network by a metric incorporating the sum of the weights on the shortest path between that node and each of the genes of interest. The weights in this network represent Bayesian posterior probabilities and range from around 3.0 to around 9.5. The shortest of these shortest paths was then taken as the appropriate relevance measure for that gene.

A simple sum over the weights on a shortest path is not an optimal approach, since a long path composed of many small weights (for example, a path of length 10, each edge of which carries a weight of 0.1) will lead to as high a weighting on the target node as a shorter path with higher weights (for example, a path of length 2, each edge of which is weighted by 0.5). Intuitively, the latter path is indicative of a much more interesting relationship than the former.

In order to avoid this issue, an exponentially discounted sum of path weights were used. The equation used for this approach is a modification of that used by Lee et al. [33] to sum weighted edges from different data sources when constructing the original yeast functional interaction network. The weights on a shortest path are ranked according to the distance from a gene of interest, and then summed using the equation

\[ W = \sum_{i=1}^{n} \frac{w_{ij}}{D^{i-1}} \] (2)

where \( n \) is the number of edges in the path, and \( w_{ij} \) is the weight on the edge between node \( i \) and node \( j \). The use of this weighting means that the weight on the edge nearest to the gene of interest is used in full, while subsequent, more distant, weights are discounted by an amount dependent upon the parameter \( D \).

The final fitness function is then, simply

\[ E(k;W) = \chi \cdot W \] (3)

that is, the coherence of a cluster multiplied by the topological relevance of the nodes it contains. Using this equation a tightly connected cluster containing only a single gene of high relevance will receive as high a score as a more loosely connected cluster all of whose nodes are highly relevant. In the search for biologically meaningful clusters of nodes with relevance to the set of nodes of interest, both types of cluster are of equal value.

The evolutionary algorithm is depicted in pseudocode in algorithm 1.

```
ALGORITHM 1. THE EVOLUTIONARY ALGORITHM

// Initialize chromosome
for i = 1 to clength
    generate random integer int in range (0, numnodes)
    parent_i = int
// Calculate fitness
f_parent = fitness(parent)
// Copy and mutate (a)
child = copy(parent)
for i = 0 to clength
    generate random double d
    if (d < mrate)
        generate random integer int in range (0, numnodes)
        child_i = int
// Calculate fitness
f_child = fitness(child)
if f_child >= f_parent
    parent = child
loop to (a) until termination criterion met
```

The genetic algorithm was run, with the parameters listed in Table 1., on the Lee network. A run ended when the maximum fitness in the population had not increased for a specified number of generations (i.e. had plateaued). At the conclusion of a run, the single fittest individual produced in the last generation of a run was taken as the cluster produced by that run. Most runs tended to converge towards multiple copies of a single individual, or minor variations thereof.

```
TABLE 1. PARAMETER VALUES FOR THE GA RUNS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial chromosome length</td>
<td>8</td>
</tr>
<tr>
<td>Population size</td>
<td>101</td>
</tr>
<tr>
<td>Number of elites</td>
<td>1</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>1 / chromosome length</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>0.4</td>
</tr>
<tr>
<td>Plateau length</td>
<td>100 gens</td>
</tr>
</tbody>
</table>
```

Authorized licensed use limited to: Newcastle University. Downloaded on May 24,2010 at 12:21:16 UTC from IEEE Xplore. Restrictions apply.
III. RESULTS

A. Network Topology and the D parameter

The evolutionary algorithm was run 100 times with a different random number seed each time, for three values of D: 1, 1.7 and 5. A value of D=1 is equivalent to a simple summation of the weights on each shortest path in a network, while higher values of D mean that, while the weight on the edges between a node of interest and its 1-neighbours and is retained in full, weights on more distant edges are successively devalued. The resulting node weight distributions for each value of D are shown in Figure 4.

![Figure 4. Distribution of weights from nodes in the LCC, a) D = 1; b) D = 1.7; c) D = 5.](image)

The genes selected by each set of runs were extracted from the entire network and visualized (Figure 5).

![Figure 5. Topology of the subnetworks selected by the algorithm. a) D = 1; b) D = 1.7; c) D = 5.](image)

The same overall pattern of clustering is apparent for all three values of D, but as D increases there is a clear trend towards selection of more individual genes, particularly more genes in two- or three-node clusters. The question of which value of the D parameter is most valuable depends to a large extent on the question being asked; two-gene interactions may be particularly interesting in a well-understood system within which a biologist is searching for previously undetected fragments of pathways, while large clusters might provide more insight into significant functional clusters.

For the purposes of this study we chose to focus on the network generated with $D = 1.7$, since it produced two large, clearly distinguishable superclusters, plus a manageable number of two- and three-node clusters. $D$ has the promise, however, to be a useful tuneable parameter, and we intend to further investigate the relationship between $D$, network topology, and the types of questions biologists usually ask.
B. Cluster Biology

In 100 runs of the algorithm with $D = 1.7$ a total of 1,145 gene instances were selected. Cluster membership was highly overlapping, however; while a significant number of genes (289; 25.2%) occurred only once, a single gene could be chosen up to 26 times (Figure 6).

In fact, the clusters formed a number of distinct superclusters, plus a large number of genes which were not attached to any other gene in the selected set (Figure 7).

There were two large, isolated clusters, together with ten clusters with three or more members, plus a number of isolated pairs of genes. The larger clusters were made up of the highly overlapping sets of genes from the right tail of the distribution in Figure 6, and naturally formed distinct superclusters.

The superclusters were annotated with terms from the Gene Ontology (GO) annotation hierarchy [47] (Summarised in Table 2). For brevity, Table 2 lists only clusters with more than two members. Individual pairs may well, however, be of interest to the biologist.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>No. genes</th>
<th>Most Common GO process</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC1</td>
<td>17</td>
<td>rRNA processing</td>
<td>4</td>
</tr>
<tr>
<td>SC2</td>
<td>26</td>
<td>protein biosynthesis</td>
<td>15</td>
</tr>
<tr>
<td>SC3</td>
<td>4</td>
<td>nuclear mRNA splicing,</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>via spliceosome</td>
<td></td>
</tr>
<tr>
<td>SC4</td>
<td>4</td>
<td>mitotic spindle</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>organization and</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>biogenesis in nucleus</td>
<td></td>
</tr>
<tr>
<td>SC5</td>
<td>4</td>
<td>ubiquitin-dependent</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>protein catabolism</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>biological process</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unknown</td>
<td></td>
</tr>
<tr>
<td>SC6</td>
<td>5</td>
<td>G1/S transition of</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mitotic cell cycle</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>G2/M transition of</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>mitotic cell cycle</td>
<td></td>
</tr>
<tr>
<td>SC7</td>
<td>3</td>
<td>telomere maintenance</td>
<td>1</td>
</tr>
<tr>
<td>SC8</td>
<td>7</td>
<td>various</td>
<td>1 each</td>
</tr>
<tr>
<td>SC9</td>
<td>4</td>
<td>various</td>
<td>1 each</td>
</tr>
<tr>
<td>SC10</td>
<td>3</td>
<td>aerobic respiration</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tricarboxylic acid cycle</td>
<td>2</td>
</tr>
<tr>
<td>SC11</td>
<td>3</td>
<td>chromatin remodelling</td>
<td>2</td>
</tr>
<tr>
<td>SC12</td>
<td>3</td>
<td>biological process</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unknown</td>
<td></td>
</tr>
</tbody>
</table>

Two major superclusters (SC1 and SC2) are prominent amongst the merged clusters (Figure 7). These clusters primarily contain proteins involved in the biogenesis (synthesis) of ribosomes (SC1) and those encoding the structural components of the ribosome (SC2). The proteins encoding the ribosomal synthetic machinery and the actual ribosomal components themselves, are well known to produce major clusters in previous studies of interaction networks [2][33], and indeed form a major component of all cellular activity [35]. Hence, the presence of SC1 and SC2 is reassuring, although not surprising.

However, included amongst these clusters are some proteins not normally associated with ribosomal structure and biogenesis, including the proteins Prs4 and YBR030W putatively involved in histidine and phospholipid metabolism and Hht1 involved in chromatin assembly and disassembly [36]. Further investigation of the reasons for the inclusion of these genes would be interesting from a biological perspective.

A number of other clusters contain proteins that are clearly functionally related. SC3 contains four proteins, all annotated as being involved in the splicing of nuclear mRNA by the spliceosome [37], SC4 contains proteins involved in mitotic spindle organization and biogenesis in the nucleus, and SC6 contains proteins thought to be involved in the G1/S and G2/M transitions of the mitotic cell cycle.

More interestingly, some clusters contain proteins of known function along with previously uncharacterised
proteins. Further investigation of the evidence used to assign edges between members of these clusters may provide insights into the function of these novel proteins. In particular, SC5 contains proteins Rpn5 and Pre6 known to be involved in ubiquitin-dependent protein catabolism associated with the proteosome [38], clustered with the proteins REH1 and FMP24, both of unknown function. SC7 contains three proteins, one of which is relevant to telomere maintenance Mak10 together with YNL181W which is a previously uncharacterised protein.

Finally, some clusters such as SC8 and SC9 contain a set of proteins that appear functionally diverse at first glance, but would clearly warrant further investigation in to the evidence associating them together, again with the possibility that proteins with multi-functional roles may be uncovered.

IV. DISCUSSION

This algorithm does appear to produce biologically plausible clusters of genes. The clusters are identified solely on the basis of network topology, but when annotated with GO ontology terms it becomes apparent that the clusters are biologically plausible, and contain genes which are potentially of relevance to the query set, involving, as they do, telomere uncapping and cell cycle genes. However, the genes of interest were selected on the basis of a common phenotype which may be achieved in a number of ways. As such, the relevance of the clusters to the process of telomere uncapping cannot be validated simply by inspection; clustering of this sort is an hypothesis generation process. Validation must be undertaken in the laboratory, and this process is currently underway.

Several of the clusters are of particular interest, particularly those containing unknown genes. Supercluster 5, for example, contains four genes, two of which are annotated as being involved in ubiquitin-dependent protein catabolism, and the other two of which are of unknown function. Ubiquitination is the process whereby proteins are tagged for degradation and is therefore fundamental to the cell cycle. The process is not fully understood at the molecular level, and the identification of new members of the pathway is potentially exciting. The evidence for this interaction is slim; inspection of the network reveals that the only evidence for the interactions of interest is co-expression in a series of microarray experiments accumulated by Lee et al. [33]. Microarray data is notoriously noisy, but given the importance of the ubiquitination process, this lead is well worth following up.

The two largest clusters are concerned with ribosomal biogenesis and structure. We have previously identified the ribosomal gene clusters using a global network clustering algorithm [19], and expressed concern that their presence may swamp the algorithm, obscuring the presence of smaller but more biologically interesting. The current algorithm clearly produces smaller, more focussed clusters. The continued presence of the ribosomal clusters could be indicative of more than just the extent of ribosomal activity in yeast. The mechanisms by which other mutations rescue cells from the effects of cdc13 may well include alterations to ribosomal function and biogenesis. Microarray screens are currently underway in DAL's laboratory to investigate this possibility.

One interesting observation is that the query genes themselves do not appear in any of the clusters. This is because they are not weighted any more highly than any other genes. Because of this, we are not looking merely for the subnetworks in the neighbourhood of the query genes (a subnetwork which could be found by other, well-established algorithms [39] and which has been taken further by Myers et al. [23]) but which pulls out clusters from anywhere in the global network.

Since we are looking for interesting interactions, we believe that eliminating singleton genes from further analysis is practical. The fact that the algorithm generates so many singleton genes is, however, interesting. It may indicate either that the algorithm could be better tuned—in particular, a stringent examination of sensitivity to parameterization may well yield better results. However, it is also possible that the singleton genes are interesting in other ways, and a closer examination of their networks neighbours which were not selected by the algorithm may be of interest.

The algorithm could be extended in a number of different ways. The possibility of weighting the query genes so as to increase the likelihood of selecting clusters containing those genes should be investigated; another parameter could be added to drive the network to either explore the neighbourhood of the genes of interest or reach further out into the network. The D parameter also affects the relative importance of nodes more distant from the genes of interest, and clearly requires further investigation.

Network clustering helps to make biological sense of large, complex biological interaction networks. Clustering algorithms such as that described here, which focus on genes of relevance to a query set of genes of interest, have the additional feature of reducing the number of genes and clusters which must be examined by biologists. The algorithm when applied to a yeast integrated functional network using a query set of genes involved in telomere capping produced biologically plausible clusters, several of which are clearly of interest to the biologists studying the cell cycle. We conclude that focussed network clustering is a valuable hypothesis generation tool for systems biology.

REFERENCES


