Identification and Evaluation of Functional Modules in Gene Co-expression Networks

Jianhua Ruan[1] and Weixiong Zhang[1,2]

[1] Department of Computer Science and Engineering,
[2] Department of Genetics,
Washington University in St. Louis
St. Louis MO 63130, USA
{jruan,zhang}@cse.wustl.edu

Abstract. Identifying gene functional modules is an important step towards elucidating gene functions at a global scale. In this paper, we introduce a simple method to construct gene co-expression networks from microarray data, and then propose an efficient spectral clustering algorithm to identify natural communities, which are relatively densely connected sub-graphs, in the network. To assess the effectiveness of our approach and its advantage over existing methods, we develop a novel method to measure the agreement between the gene communities and the modular structures in other reference networks, including protein-protein interaction networks, transcriptional regulatory networks, and gene networks derived from gene annotations. We evaluate the proposed methods on two large-scale gene expression data in budding yeast and Arabidopsis thaliana. The results show that the clusters identified by our method are functionally more coherent than the clusters from several standard clustering algorithms, such as k-means, self-organizing maps, and spectral clustering, and have high agreement to the modular structures in the reference networks.

Key words: clustering, community identification, microarray, co-expression networks

1 Introduction

Many biological sub-systems considered in systems biology can be modeled as networks, where nodes are entities such as genes or proteins, and edges are the relationships between pairs of entities. Examples of biological networks include protein-protein interaction (PPI) networks [1], gene co-expression networks [2], metabolic networks [3], and transcriptional regulatory networks [4]. Much effort has been devoted to the study of their overall topological properties and similarities to other real-world networks [5–8].

A large amount of available gene expression microarray data has provided opportunities for studying gene functions on a global scale. Since genes that are on the same pathways or in the same functional complex are often regulated by the same transcription factors (TFs), they usually exhibit similar expression patterns under diverse temporal and physiological conditions. Therefore, an important step in analyzing gene functions is to cluster genes according to their expression patterns. The clusters can then be analyzed in several ways. For example, from the promoter sequences of the genes in the same cluster, one may identify common short DNA sequences, which can often suggest the regulation pathways of the genes; in addition, if the majority of the genes in a cluster are known to have some common functions, it is likely that the unannotated genes in the same cluster may also share similar functions. (See [9] for a review). The most popular clustering techniques for gene expression data include hierarchical clustering [10], k-means clustering [11], and self-organizing maps (SOM) [12].

However, genes of similar expression patterns may not necessarily have the same or similar functions. Genes could be accidentally co-regulated or co-expressed [2]; a single event often activate multiple pathways that have distinct biological functions. On the other hand, genes with related functions may not show any close correlation in their expression patterns. For example, there might be time-shift between the expression patterns of genes in the same pathway [13]. Most existing clustering algorithms do not take these possibilities into account.

Another challenging problem for clustering algorithms is to determine the most appropriate number of clusters without prior knowledge of the data. For most clustering algorithms, such as k-means and SOM, it is the user's responsibility to decide the number of clusters to be computed, and it is always possible for the algorithms to return the specified numbers of clusters, regardless of the structure of the data.

To objectively evaluate and validate clustering results is also a daunting task. Generally, different clustering algorithms provide different results and unveil different aspects of the data. To assess the quality of clustering results, most studies have focused on the separation between clusters or homogeneity within clusters [14]. Such numerical evaluation methods depend solely on the data and face a common dilemma: one cannot maximize both the separation and homogeneity at the same time. More importantly, these methods seldomly perform any reality check. For example, does a clustering make any biological sense? Several alternative approaches have been proposed to validate clustering results with biological knowledge, for example, using annotations in the gene ontology (GO) [15, 16]. However, these methods are usually affected by factors such as the number of clusters and the distribution of cluster sizes, and cannot precisely measure clustering qualities.

Here, we take a network-based perspective to efficiently *identify* and *evaluate* intrinsic modular structures embedded in large gene expression data. Given the expression profiles of a set of genes, we first construct a co-expression (CoE) network, where the nodes in the network are genes, and the edges reflect expression similarities between pairs of genes. We then apply an algorithm that we have developed recently to identify natural communities in the network, which are densely connected subgraphs that are unexpected by chance [17, 18]. Compared to existing clustering methods, our algorithm is relatively independent of any detailed domain knowledge, and can automatically determine the best number of clusters based on the internal structure of the data. Furthermore, we also propose a method to evaluate the biological significance of the clustering results based on their agreement with the structure of other reference biological networks.

We apply the methods to two large gene expression datasets, one for yeast and the other for Arabidopsis. We evaluate the clustering results on yeast genes with three reference networks, including a protein-protein interaction (PPI) network [19], a network based on GO annotations [20], and a network based on TF biding data measured with ChIP-chip technology [21], and the results on Arabidopsis genes with a GO-based reference network. We compare our results with several popular clustering algorithms, including k-means, SOM and spectral clustering, which are applied directly to the expression data. The comparison shows that our network-based approach discovers significantly more enriched functional groups, which also have a better agreement with the reference networks.

The paper is organized as follows. In section 2, we describe the method for constructing gene CoE networks, the algorithm for community identification, and the approach for cluster evaluation. In section 3, we first present some topological results of the CoE networks, then discuss our clustering results and compare them with the results from several popular clustering algorithms. We conclude in section 4 with some discussion.

2 Methods

2.1 Constructing Gene CoE Networks

Many methods have been proposed for constructing CoE networks from gene expression data. The most popular methods first compute a similarity between the expression profiles of every pair of genes, and determine a threshold to select pairs of genes to be connected [22–24]. The problem with this type of approaches, aside from being arbitrary in choosing a threshold, is that gene CoE often exhibits a local-scaling property. For example, genes in one cluster may be highly correlated to one another, while genes in another group may be only loosely correlated. Therefore, if we choose a stringent threshold value, many genes in a loosely correlated group may become unconnected. On the other hand, if we attempt to include more gene in the network, the threshold may have to be so low that a large portion of genes are almost completely connected, making further analysis a difficult task. For example, to construct a CoE network for the 3000 yeast genes that we will see in Section 3.1, even if we allow 10% of the genes to be unconnected, the majority of the genes still have more than 300 links (Fig. 1).

We propose a rank-based transformation of similarity matrices to deal with such local-scaling property. We first calculate the Pearson correlation coefficient (or some other similarity measures) between every pair of genes. Then for every gene, we rank all other genes by their correlation coefficients to the gene. Given the ranks, we connect every gene to its top α co-expressed genes, where α is a

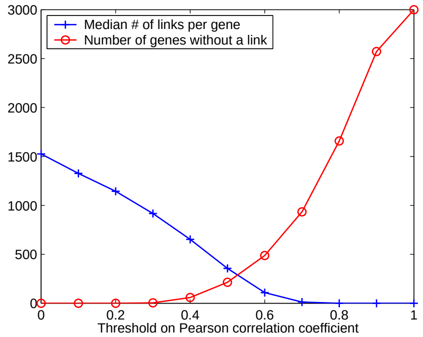

Fig. 1. Median number of CoE links per gene and the number of genes without a CoE link as a function of the threshold on the Pearson correlation coefficient.

user defined threshold, with values typically smaller than 5. Note that although the correlation coefficient matrix, C, is symmetric, i.e. $C(i, j) = C(j, i)$, the rank of gene i with respect to gene j, $R(i, j)$, is in general not equal to the rank of gene j with respect to gene i, $R(j, i)$.

This network has several important features. First, all nodes are connected, since each node is connected to at least α other nodes. By varying α, we obtain networks of different granularities. Second, some nodes may have more than α edges, due to the asymmetric property of the ranking. That is, although gene A lists only α genes as its friends, other genes that are not in A's friend list may have A as their friends. In other words, the network can be viewed as directed, even though the directions are ignored in our clustering. In section 3.1, we will show that a CoE network thus constructed has a prominent topological feature different from the CoE networks obtained in previous studies [2, 24, 25].

A network constructed with this procedure may be different from the underlying biological network that regulates the genes. Nevertheless, at a higher level, the network may capture some topological properties of the actual regulatory network and preserve functional relationships among genes. Genes that are in the same pathway or functional complex tend to be close to one another in the network, i.e., they are often directly linked to each other or connected by short paths. As we will see in section 3, clustering of such networks can indeed produce biologically more meaningful modules than clustering the original expression data with a conventional clustering method. We will also show that clustering of this network is rather robust, in that perturbing a large fraction of its connections does not significantly affect the final clustering results.

2.2 Community Identification

Identifying community structures in a network is similar, but not equivalent, to the conventional graph partitioning problem; both amount to clustering vertices into densely connected subgraphs [26]. A key difference is that for the former, we need to decide whether there are indeed natural communities and how many communities exist in a given network. In contrast, in conventional graph partitioning, the user has to decide how many clusters to look for.

We recently proposed a spectral-based community identification method [17, 18]. The method has several unique features. First, it considers local neighborhood information of each node to improves clustering quality [17]. Second, the algorithm combines a modularity function Q to automatically determine the most appropriate number of clusters in a network. Third, the algorithm can handle networks of several thousands of nodes in a few minutes, much faster than most existing algorithms, while often achieving better clustering qualities. We have extensively tested the algorithm on many simulated networks and real-world networks with known community structures, as well as several real applications such as PPI networks and scientific collaboration networks. The results from these analyses show that our method is both efficient and effective. The detailed analysis and evaluation of the algorithm can be found in [18]. Here we briefly describe the key ideas in the algorithm.

Modularity Function To determine the optimal community structure of a network, Newman and Girvan [27] recently proposed a modularity function, Q, which is defined as:

$$Q(\mathit{\Gamma}_k) = \sum_{i=1}^{k}(e_{ii} - a_i^2), \tag{1}$$

where $\mathit{\Gamma}_k$ is a clustering that partitions the nodes in a graph into k groups, e_{ii} is the fraction of edges with both nodes within cluster i, and a_i is the fraction of edges with one or both nodes in cluster i. Intuitively, the Q function measures the percentage of edges fully contained within the clusters, subtracted by what one would expect if the edges were randomly placed. The value of Q is between -1 and 1; a larger Q value means stronger community structures. If a partition gives no more within-cluster edges than expected by chance, $Q \leq 0$. For a trivial partitioning with a single cluster, $Q = 0$. It has been observed that most real-world networks have $Q > 0.3$ [28]. The Q function can also be extended to weighted networks straightforwardly by generalizing e_{ii} and a_i to edge weights, instead of number of edges.

It has been shown empirically that higher Q values correspond to better clusters in general [27, 29]. Therefore, the Q function provides a good quality measure to compare different community structures, and can serve as an objective function to search for the optimal clustering of a network.

The *Qcut* Algorithm Several clustering algorithms have been developed based on *approximate* optimization of Q (as surveyed in [29]), since the optimization is NP-hard [30]. Among them, a spectral algorithm NJW [31], can approximately optimize Q if the number of clusters (k) is given, as shown in [32]. To automatically determine the number of clusters, the NJW algorithm is executed multiple times, with k ranging from the user defined minimum K_{min} to maximum K_{max}

number of clusters. The k that gives the highest Q value is deemed the most appropriate number of clusters. The idea has been implemented recently by others and us [32, 17].

While this idea is effective in finding community structures in small networks, it scales poorly to large networks, because it needs to execute NJW, whose running time is $O(n^2)$, up to K_{max} times. Without any prior knowledge of a network, one may over-estimate K_{max} in order to reach the optimal Q. In the worst case, K_{max} can be linear in the number of vertices, making it impractical to iterate over all possible k's for large networks.

In order to develop a method that scales well to large networks while retaining effectiveness in finding good communities, we developed an algorithm, called *Qcut*, to recursively divide a network into smaller ones while optimizing Q [18].

Given the adjacency matrix of a network G, we apply the standard NJW spectral clustering algorithm [31] to search for an up to l-way partitioning, where l is a small integer ($l < 5$ typically), that gives the highest Q value. Then, the algorithm is recursively applied to each subnetwork, until the overall Q value cannot be improved by any further partitioning. At each step, a (sub)network is divided into k subnetworks, where k is between 2 and l, and can be different for each partitioning. To reduce computation cost, we restrict l to small integers. We find that with l as small as 3 or 4, the *Qcut* algorithm can significantly improve the Q values over standard two-way partitioning strategies [33, 32], and is much more efficient than direct k-way methods [32, 17]. After each split and at the end of all splits, an efficient procedure is applied to fine-tune the clusters in order to further improve the modularity, making *Qcut* one of the most effective (in terms of accuracy) and efficient algorithms in community identification.

2.3 Cluster Evaluation

A conventional way for evaluating clustering results is to measure separation and homogeneity. We are more interested in the biological soundness and relevance of the clustering results. Therefore, we use two methods based on gene functional annotations to evaluate clustering qualities obtained from gene CoE networks.

Statistical Enrichment of GO Terms To assess the functional significance of gene clusters, we first compute the enrichment of GO terms for the genes within each cluster. The statistical significance of GO term enrichment is measured by a cumulative hypergeometric test [34]. The p-values are adjusted by Bonferroni corrections for multiple tests [34]. The search of enriched GO terms is performed with a computer program GO::TermFinder [35].

To compare different clustering results, we count the number of GO terms enriched in the clusters at a given significance level. Furthermore, to rule out the possibility that a single cluster may contain a very large number of enriched GO terms and therefore dominate the contribution from other clusters, we also count the number of clusters that have at least one enriched GO term at a given significance level. Note that two clustering results cannot be compared by this method if they differ significantly in numbers of clusters or cluster size distributions, which may strongly affect the number of enriched GO terms. The results of the comparison also depend on what p-value threshold is used.

Evaluation Using Reference Networks We propose a novel method for assessing clustering qualities based on external information of the genes. The basic idea is to introduce a functional reference network (discussed later), and compare the clustering of the CoE network with the reference networks. In such a reference network, genes are linked by edges that represent certain functional relationships between them, where the edges may be weighted according to the reliability or significance of the relationships. This network can be expected to have some modular structures as well. Since our purpose is to identify functional modules within a CoE network, we would prefer a good clustering of the CoE network to represent a good partitioning of the reference network as well; i.e., genes within the same CoE clusters should be connected by many high weight edges in the reference network, while genes in different CoE clusters should share less functions or be connected with low weight edges in the reference network. To measure the agreement between the clustering of a CoE network and a reference network, we force the reference network to be partitioned exactly the same way as the CoE network, i.e., the group memberships of the nodes in the reference network are the same as that of the CoE network. We then compute the modularity of the reference network using Equation (1). Since the modularity score is not biased by the number of clusters or the cluster size distributions, it can be applied to compare arbitrary clustering results.

Now that we have introduced the measurement, what can be a reference network and how do we get it? First, many available biological networks, such as PPI networks and genetic interaction networks, can be adopted directly. Evidently, however, some networks may be more suitable than others for evaluating gene CoE clusters.

In general, a reference network does not have to be directly observed from experiments, but rather derived from knowledge about the genes. Two genes can be connected if they posses some common attributes or features, given that the common attributes are related to CoE. For example, they may participate in the same biological process or be regulated by a common TF. These types of information can be represented by a matrix, where each row is a gene, and each column is an attribute. To construct a network from the matrix, genes are treated as nodes, and an edge is drawn between two genes if they share at least one common attribute. Edges are weighted by some similarity measure of genes' attributes. To measure the similarity, we use a well-developed function in document clustering that takes into account the significance of attributes [36]. For example, the GO terms GO:0009987 (cellular process), which is very close to the root of the GO graph and has a large number of genes associated, is not very informative in clustering genes and should be weighted less than the GO term GO:0045911 (positive regulation of DNA recombination).

Denote a gene-attribute matrix by $A = (a_{ij})$, where $a_{ij} = 1$ if gene i has attribute j, or 0 otherwise. A is transformed into a weighted matrix $W = (w_{ij})$, where $w_{ij} = a_{ij} \times idf_j$. The weighting factor idf_j, called the inverse document frequency (IDF) [36], is defined by $idf_j = \log(n/\sum_i a_{ij})$, where n is the number of genes. With this transformation, the attributes that occur in many genes receive low weights in W. The edge weight between two genes is then measured by the cosine of their weighted attribute vectors:

$$S_{ij} = cos(w_{i.}, w_{j.}) = \frac{\sum_k w_{ik} w_{jk}}{\sqrt{\sum_k w_{ik}^2 \sum_k w_{jk}^2}}, \tag{2}$$

where $w_{i.}$ and $w_{j.}$ are the i-th and j-th rows of W, respectively. As expected, many genes may be connected with very low weights if they share some non-specific functions. We apply a weight cutoff to remove such edges. We have found, however, that the result is almost not affected by the use of different cutoff values, as discussed in Results section.

We use three types of reference networks to evaluate clusters. The first is a network constructed from biological process GO annotations [20], with each term as an attribute. The ontology and annotation files for yeast and Arabidopsis genes are downloaded from `http://www.geneontology.org/`. To construct a reference network, we first convert the original annotation files to include complete annotations, i.e., if a gene is associated with a certain term, we also add all ancestors of the term into the gene's attribute list due to term inheritance. If two terms are associated with exactly the same set of genes, we remove one to avoid double counting. We also remove GO terms that are associated with more than 500 or less than 5 genes. The procedure results in 1034 and 438 GO terms for yeast and Arabidopsis, respectively. The second is a PPI network for budding yeast, downloaded from the BioGRID database [19]. We combined all physical interactions obtained from yeast two-hybrid or affinity purification-mass spectrometry experiments. The edges are weighted by the number of times an interaction was observed. The third network is a co-binding network derived from the ChIP-chip data of 203 yeast transcription factors (TFs) under rich media conditions [21]. We treat each TF as an attribute, and construct a network with the procedure described above. We only consider a binding as genuine if its p-value is less than 0.001, according to the original authors [21].

3 Results

3.1 Topology of Yeast CoE Networks

Previous studies have analyzed the topologies of various networks, including biological and social networks, and suggested a common scale-free property [5–8]. In a scale-free network, the probability for a node to have n edges obeys a power-law distribution, i.e. $P(n) = c \times n^{-\gamma}$. The implication of the scale-free property is that a few nodes in the network are highly connected, acting as

Table 1. Statistics of yeast CoE networks.

α	2	3	4	5
# of nodes	3000	3000	3000	3000
# of edges	5432	8103	10775	13432
k_{avg}	3.6	5.4	7.2	9.0
c	0.089	0.124	0.144	0.159
c_r	0.010	0.015	0.018	0.02
c_{sf}	0.002	0.003	0.004	0.005

k_{avg}: averge node degree; c: clustering coefficient; c_r: clustering coefficient of the network constructed from permuted expression data; c_{sf}: clustering coefficient of the rewired network.

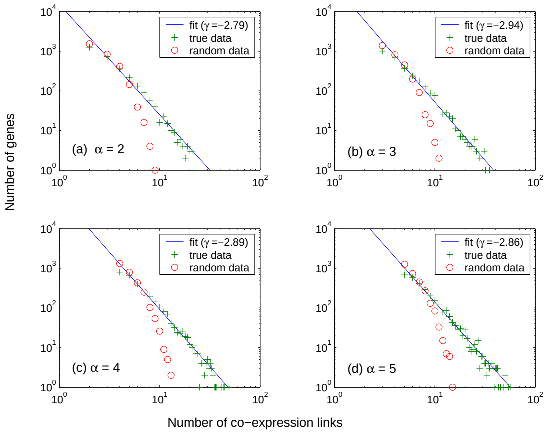

Fig. 2. Distribution of the number of CoE links. Y-axes show the number of genes with a certain number of CoE links (X-axes) in a network.

hubs, while most nodes have low degrees. In contrast, in a random network, connections are spread almost uniformly across all nodes. Real networks also differ from random networks in that the former often have stronger modular structures, reflected by higher clustering coefficients [28].

In this study, we obtained a set of yeast gene expression data measured in 173 different time points under various stress conditions [37], and selected 3000 genes that showed the most expression variations. We constructed four CoE networks with α = 2, 3, 4 and 5, respectively, i.e., we let each gene connect to its top α correlated genes (see section 2.1). To compare, we also randomly shuffled the real gene expression data, and constructed four networks from the random data with the same α values.

To determine the topological characteristics of the CoE networks, we first plotted the number of genes having n connections as a function of n in a log-log scale. As shown in Fig. 2, the networks constructed from the real data exhibit

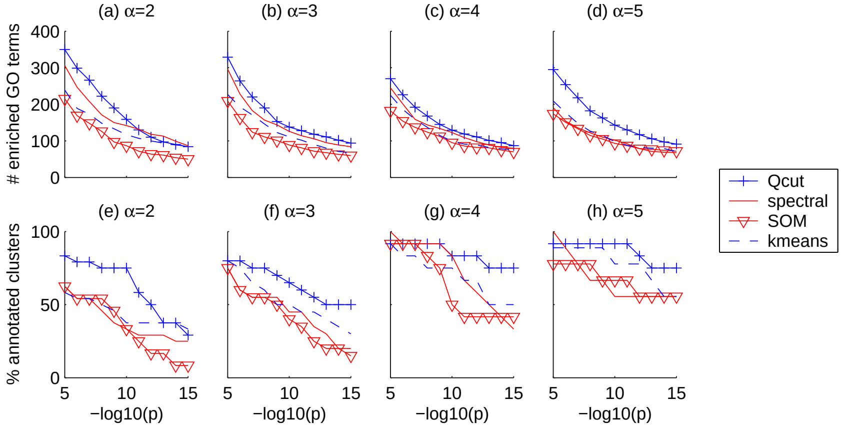

Fig. 3. Enrichment of GO terms in yeast CoE networks. Y-axes in (a)-(d): number of GO terms enriched in the clusters. Y-axes in (e)-(h): percentage of clusters that are enriched with at least one GO term. X-axes: p-value cutoff to consider a GO term enriched.

a power-law degree distribution for all the α values considered, indicating that an overall scale-free topology is a fairly robust feature of the CoE networks. In contrast, the networks constructed from the randomized expression data are close to random networks and contain significantly fewer high-degree nodes. Second, we computed the clustering coefficients of the networks derived from real and randomized expression data. As shown in Table 1, the true CoE networks have much higher clustering coefficients than the random network. Furthermore, we permuted the CoE networks through random rewiring [38], which preserves degree for each node, and thus does not change the scale-free property of the networks. As shown in Table 1, the clustering coefficients of the rewired networks are significantly lower than that of the original networks, indicating that high clustering coefficients is indeed a property of CoE networks.

It is not surprising to see that CoE network is yet another example of scale-free networks. However, several previous studies on a number of gene CoE networks have suggested that there might exist profound topological differences between gene CoE networks and other biological networks [2, 23, 25]. In these studies, it has been observed that the exponent γ for the power law degree distribution of CoE networks is consistently less than 2, while in other biological networks, including PPI networks and metabolic networks, as well as in real-world social and technology networks, γ is usually between 2 and 3 (for examples see [28, 38]). A scale-free network with $\gamma < 2$ has no finite mean degree when its size grows to infinity, and is dominated by nodes with large degrees [28]. To determine the values of γ for the CoE networks that we have constructed, we fitted a linear regression model to each log-log plot to calculate its slope. As

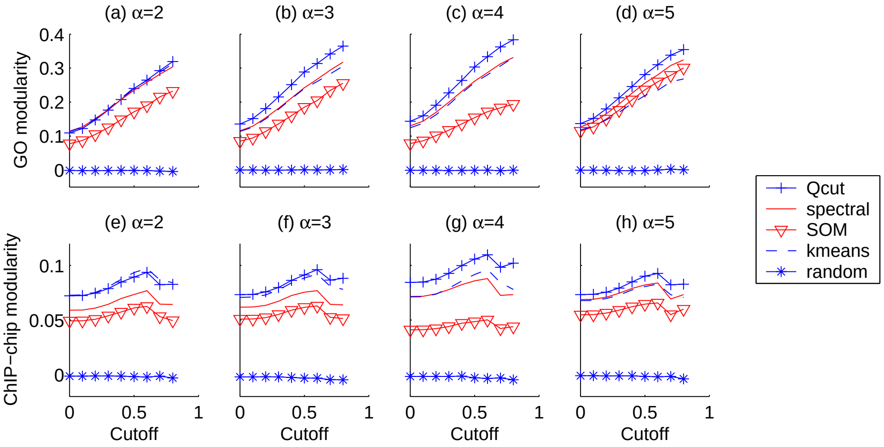

Fig. 4. Agreement between modular structures in yeast CoE networks and two reference networks derived from GO annotations (a-d) and ChIP-chip data (e-h). X-axes: edge weight cutoff for the reference networks.

shown in Fig. 2, the values of γ in our networks are consistently between 2 and 3, similar to many real-world or biological networks.

The difference in γ between previous CoE networks and ours is most likely due to the difference in the network construction procedures. We used a rank-based method in selecting CoE links, while most existing methods are threshold-based. A threshold-based network tends to include a large number of high degree nodes, and therefore usually have a small γ value. Although further work is required, the similarity in γ values between our networks and other biological and real-world networks suggests that the networks constructed by our method may better represent the underlying functional structures than previous CoE networks.

3.2 Functional Modules in Yeast CoE Networks

We applied the *Qcut* algorithm to cluster the four CoE networks constructed in section 3.1. The best numbers of clusters suggested by *Qcut* for the four networks are 24, 20, 12 and 12, respectively. For comparison, we also applied three popular clustering algorithms, including k-means, SOM, and spectral clustering, to the expression data, using Pearson correlation-coefficient as the distance measure. We obtained k = 24, 20, 12 and 9 clusters for each of the three competing algorithms. The SOM algorithm was executed on 4 × 6, 4 × 5, 3 × 4, and 3 × 3 grids to produce the desired number of clusters [12]. Because *Qcut* identified 12 clusters on both the $\alpha = 4$ and $\alpha = 5$ networks, we matched the 12 clusters of the $\alpha = 5$ network with the 9 clusters from the competing algorithms to avoid redundant comparison. Another reason for this matching is that *Qcut* often produce a few small clusters, while the clusters of the competing algorithms are relatively uniform in sizes. Therefore, the “effective” number of clusters is smaller for *Qcut* than for other algorithms, so we used the last test to compensate some differences in the cluster size distributions.

To validate the biological significance of the clusters, we first counted the number of GO terms enriched in the clusters and the number of clusters that had at least one enriched GO term at various significance levels. As shown in Fig. 3, the clusters identified by *Qcut* contain more enriched GO terms than the competing algorithms for most p-value cutoff levels and for different number of clusters (Fig. 3(a)-(d)). Furthermore, the percentages of clusters containing at least one enriched GO term are also higher for *Qcut* than for the other algorithms (Fig. 3(e)-(h)). However, as observable from the figure, the number of enriched GO terms increase with the number of clusters. Therefore, it is hard to conclude which network has produced the best clustering result.

Second, we evaluated the clusters with three reference networks that capture different functional interactions between genes: a co-function network based on GO annotations, a co-binding network based on ChIP-chip data, and a PPI network (Section 2.3).

The comparison with all three reference networks indicates that the clusters identified by *Qcut* have higher agreement with the reference networks than do the clusters by the competing algorithms (Fig. 4 and 5). The spectral clustering algorithm generally performs better than the other two, which is reasonable since the spectral method is able to capture some topological features embedded in the data. We also randomly shuffled the clustering results of *Qcut* while fixing the sizes of the clusters, and compared the random clusters with the three reference networks. The modularity is always very close to zero (Fig. 4 and 5), meaning that the agreement between our clustering results and the three reference networks is not due to chance.

Among the three reference networks, the GO-based network has higher agreement with the CoE network modules ($Q > 0.35$) than do the PPI network ($Q \approx 0.15$) and the co-binding network ($Q \approx 0.1$). The low agreement between CoE and PPI networks may be partially due to the high level of noises in PPI data. On the other hand, the low agreement between the CoE and co-binding networks is somewhat unexpected, because co-binding should be a relatively strong evidence of CoE. The reason might be that the gene expression data were measured under stress conditions while the ChIP-chip experiments were conducted under normal conditions. Therefore, genes bound by common TFs under normal conditions may not necessarily exhibit similar expression patterns under these stress conditions, and some co-binding under stress conditions were not captured by the ChIP-chip experiments.

For the GO-based reference network, the modularity value is a monotonic increasing function of edge cutoffs, indicating that genes sharing many functions or several specific functional terms are more likely to be co-expressed than genes sharing some broad functional terms. In comparison, the ChIP-chip modularity reaches its peak at cutoff = 0.6, probably because there are relatively fewer genes sharing exactly the same regulators, and therefore the co-binding network

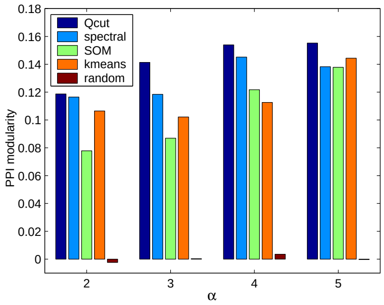

Fig. 5. Agreement between modular structures in yeast CoE networks and a PPI network. The four groups represent four networks with different values of α.

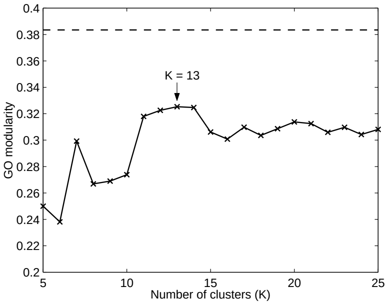

Fig. 6. Agreement between the clusters identified by spectral clustering and the GO-based reference network as a function of the number of clusters (k). The dashed line represents the best agreement achieved by *Qcut* (α=4, k=12).

becomes very sparse when the cutoff is greater than 0.6. However, the relative performance of different clustering algorithms is not affected by the cutoffs.

Both Fig. 4 and Fig. 5 show that the modules in the $\alpha = 2$ CoE network have the worst agreement with any of the three reference networks, which means that this network might be too sparse to capture all functional relationships. The $\alpha = 4$ CoE network has the highest agreement with the three reference networks, while the networks with $\alpha = 3$ or 5 give slightly worse results.

Furthermore, to test if the competing algorithms may give the best results with a different number of clusters, we applied the spectral clustering to obtain $k = 5, 6, \ldots, 25$ clusters, and computed their agreement to the GO-based reference network at cutoff value = 0.8. As shown in Fig. 6, the spectral clustering achieved best modularity 0.323 at $k = 13$, which is significantly lower than the best modularity of *Qcut* ($Q = 0.384$).

Finally, Table 2 shows the number of genes within each cluster identified from the $\alpha = 4$ network, the most significantly enriched GO biological process terms, and the transcription factors that may bind to the genes within each cluster. As shown, most clusters contain highly coherent functional groups, and are regulated by a few common transcription factors, e.g., clusters 8, 9, 11 and 12. The majority of the genes in cluster 12 are involved in protein biosynthesis, and can be bound by FHL1 and RAP1, both of which are known to be involved in rRNA processing and regulating ribosomal proteins [39]. Cluster 9 is significantly enriched by genes that are involved in generation of precursor metabolites and energy, and can be bound by HAP4, a TF regulating carbohydrate metabolism [39]. Cluster 2 contains almost two third of the ribosome biogenesis genes, although no TFs bind to this set of genes specifically. Cluster 11 are enriched with genes that can be bound by eight different TFs. Interestingly, these TFs are all known cell-cycle regulators [39].

Several small clusters correspond to very specific functional groups. For example, 17 of 22 genes in cluster 10 are involved in Ty element transposition; 9 of 18 genes in cluster 3 are related to chromatin assembly or disassembly. Six genes in cluster 3 are regulated by HIR1/2/3, which are known to be involved in the transcription of histone genes [39].

Among the 25 genes in cluster 4, 4 genes have a common function in telomere maintenance, while 16 genes encode hypothetic proteins and have unknown functions. Interestingly, 5 of the 16 uncharacterized genes are located near telomeric region [39]. Moreover, A significant number of genes in this cluster are regulated by four common transcription factors (Table 2). Therefore, it is very likely that these uncharacterized genes are closely related to the function or maintenance of telomere. Clusters 5 and 7 contain both a large fraction of genes with unknown functions, and groups of genes with significantly enriched common functions or common TFs. It is possible that these uncharacterized genes also have similar functions to the other annotated genes in the same cluster.

3.3 Robustness of Clustering Results

Since gene expression measurement is inherently noisy, and our method only used the top-ranked CoE edges in network construction, we need to evaluate whether the resulting clusters were stable with respect to perturbations. To this end, we removed all the top three CoE links from the yeast $\alpha = 6$ network. That is, each gene was connected only to its fourth, fifth and sixth best correlated genes. This network has about the same number of edges as the $\alpha = 3$ network, but very different edges. In fact, the edges in the two networks are completely different. To compare their modular structures, we calculated a minimal Wallace Index [40] between the clustering results on the two networks, which is a defined by $W(\Gamma, \Gamma') = \min\left(N_{11}/S(\Gamma), N_{11}/S(\Gamma')\right)$, where Γ and Γ' are two clustering results for comparison, N_{11} is the number of pairs of genes in the same cluster in both Γ and Γ', and $S(\Gamma)$ is the number of pairs of genes in the same cluster in Γ.

Surprisingly, the clustering on these two network are fairly similar: the Wallace Index between the two clusters is 0.63, i.e., 63% of the gene pairs are conserved between the two clustering results. In contrast, we would only expect the two clusters to share (12±0.1)% of the gene pairs if the two networks were not related. Furthermore, the clusters obtained from the reduced $\alpha = 6$ network

Table 2. Functional modules in a yeast CoE network.

Cluster	Size	Category[1]	Term	Count	Enrichment[2]	P-value
1	361	BP	protein catabolism	32	4.2	2.0E-12
		BP	protein folding	21	5.9	1.6E-11
2	498	BP	ribosome biogenesis	133	9.2	2.0E-106
3	18	BP	chromatin assembly or disassembly	9	36.4	5.3E-13
		TF	HIR2	6	129.8	2.3E-12
		TF	HIR1	6	62.9	3.0E-10
		TF	HIR3	6	57.7	5.3E-10
4	25	BP	telomerase-independent telomere maintenance	4	82.3	1.1E-07
		BP	biological process unknown	16	2.9	7.6E-06
		TF	GAT3	13	56.8	3.5E-21
		TF	YAP5	15	43.5	5.8E-17
		TF	PDR1	9	25.8	3.1E-11
		TF	MSN4	8	35.0	3.8E-11
5	422	BP	spore wall assembly	16	7.0	1.6E-10
		BP	biological process unknown	138	1.5	1.2E-07
		TF	NRG1	21	4.2	1.4E-08
		TF	SUM1	16	3.9	2.3E-06
		TF	PHD1	15	3.4	3.2E-05
6	99	–	–	–	–	–
7	463	BP	carbohydrate metabolism	41	2.9	4.6E-10
		BP	biological process unknown	178	1.7	9.5E-17
		BP	response to stimulus	62	1.7	2.0E-05
		TF	UME6	25	2.5	2.6E-05
		TF	NRG1	15	2.8	3.6E-04
8	108	BP	nitrogen compound metabolism	25	7.0	5.2E-15
		TF	MET31	4	9.6	8.0E-04
		TF	MET32	5	5.7	2.1E-03
9	192	BP	generation of precursor metabolites and energy	50	8.2	7.5E-33
		TF	HAP4	22	9.2	5.1E-16
10	22	BP	Ty element transposition	17	58.6	6.2E-29
		TF	SUM1	4	18.9	5.8E-05
11	604	BP	carboxylic acid metabolism	76	3.0	2.4E-19
		BP	cell organization and biogenesis	212	1.6	3.7E-15
		TF	SWI6	45	2.9	3.6E-11
		TF	SWI4	44	2.8	2.7E-10
		TF	FKH2	35	3.0	4.7E-09
		TF	MBP1	36	2.8	1.9E-08
		TF	STE12	22	3.6	7.9E-08
		TF	NDD1	30	2.9	1.1E-07
		TF	FKH1	34	2.5	9.6E-07
		TF	MCM1	22	2.9	3.9E-06
12	186	BP	protein biosynthesis	131	6.4	6.4E-85
		TF	FHL1	96	17.1	3.3E-105
		TF	RAP1	58	11.5	2.2E-48

[1] For each cluster, significantly enriched biological process GO terms (BP) or binding of transcription factors (TF) are counted.

[2] Fold of enrichment is calcuated as:

$$\frac{\text{(number of genes in cluster with the term)} \times \text{(number of genes in genome)}}{\text{(number of genes in cluster)} \times \text{(number of genes in genome with the term)}}.$$

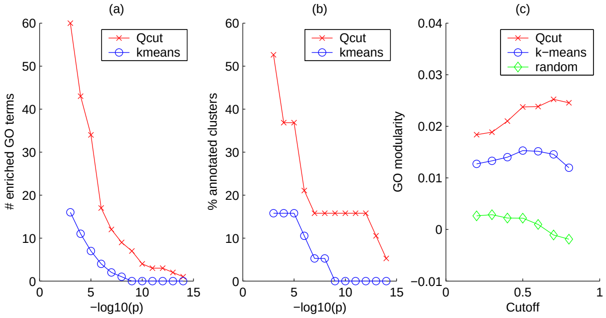

Fig. 7. Enrichment of GO terms in the Arabidopsis CoE network. (a) number of enriched GO terms; (b) percentage of clusters with at least one enriched GO term; (c) agreement between modular structures in the Arabidopsis CoE network and a reference network derived from GO annotations. X-axes in (a) and (b) are p-value cutoff to consider a GO term enriched. X-axis in (c) is edge weight cutoff for the reference network.

still contain significantly more enriched GO terms than the clusters identified by k-means and SOM (data not shown).

3.4 Functional Modules in an Arabidopsis CoE Network

To test our method on higher organisms, we applied it to a set of Arabidopsis gene expression data downloaded from the AtGenExpress database (`http://www.uni-tuebingen.de/plantphys/AFGN/atgenex.htm`). The dataset contains the expression of 22k Arabidopsis genes in root and shoot in 6 time points following cold stress treatment. We selected the genes that are up- or down-regulated by at least five-folds in at least one of the 6 time points in root or shoot. We then constructed a CoE network by connecting each gene to its top three correlated genes (i.e. $\alpha = 3$). The network has 2545 genes and 5838 CoE links.

Our clustering algorithm partitioned the network into 19 clusters, with a Q value of 0.81, indicating strong modular structures. As in the previous experiments, we examined the GO terms enriched in the clusters at various significance levels, and compared them with the results of the standard k-means algorithm that partitions the gene expression data into 19 clusters. As shown in Fig. 7(a) and (b), the clusters identified by our network-based method contains significantly more enriched terms than that identified by the k-means, and GO terms are enriched in more clusters in our method than in k-means. Furthermore, the comparison with a reference network derived from GO annotations (section 2.3) shows that the clusters identified by *Qcut* is more consistent with the reference network (Fig 7(c)). Note that due to the high complexity of gene expression regulation and the lack of detailed gene annotations, the modularity of the GO network in Arabidopsis is much lower than that of yeast (0.025 vs 0.38).

Table 3. Functional modules in an Arabidopsis CoE network.

Cluster	Size	GO term	Count	Enrichment*	P-value
1	199	-	-	-	-
2	141	-	-	-	-
3	79	-	-	-	-
4	180	catalytic activity	99	1.6	4.1E-09
		amino acid and derivative metabolism	18	4.4	3.9E-08
5	284	endomembrane system	79	1.6	3.5E-06
6	238	oxidoreductase activity	40	2.6	7.7E-09
		secondary metabolism	18	3.1	9.9E-06
7	65	photosynthesis	11	32.6	8.7E-16
8	261	RNA binding	11	4.6	9.2E-06
9	186	galactolipid biosynthesis	3	17.6	1.8E-04
10	19	branched-chain-amino-acid transaminase activity	3	172.6	1.7E-07
11	117	starch metabolism	4	16.0	5.0E-05
		circadian rhythm	6	7.6	8.5E-05
12	271	protein modification	37	2.1	4.3E-06
13	268	methyltransferase activity	8	4.7	1.4E-04
14	13	response to heat	8	87.7	1.9E-15
15	223	antiporter activity	10	6.1	1.5E-06
16	151	transcription regulator activity	60	3.0	2.5E-17
17	200	zeaxanthin epoxidase activity	3	16.4	2.2E-04
18	17	lipid binding	5	48.2	2.9E-08
		membrane	12	2.7	1.8E-04
19	249	calcium ion binding	13	3.2	1.1E-04

*See Table 2

Table 3 shows the most enriched functional categories for each cluster. Some clusters are enriched with functions that are known to be related to cold stress responses, e.g. clusters 7 (photosynthesis), 11 (circadian rhythm), 14 (response to heat), 15 (antiporter activity) and 18 (lipid binding). Since the annotation for the Arabidopsis genome is much poorer than that for the yeast genome, the enrichment of GO terms in the clusters for Arabidopsis genes are not as significant as that for the yeast genes. On the other hand, our method may be applied to assign putative functional roles to some of these unannotated genes.

4 Conclusions and Discussion

In this paper, we proposed a network-based method for clustering microarray gene expression data, and a method for evaluating clustering results based on reference networks. We introduced a simple rank-based method to construct gene CoE networks from microarray data, and applied a spectral clustering algorithm that we developed recently to cluster networks into densely connected sub-graphs. We applied our method to two gene expression datasets, and showed that the network-based clustering method can produce biologically more meaningful clusters than conventional methods such as k-means and SOM. The clusters identified by our methods contain significantly more enriched GO terms than other algorithms and exhibited better agreement with several reference networks.

It is rather surprising that the simple method we proposed to construct CoE networks worked well. The connections in such a CoE network are obviously different from actual biological interactions. Nevertheless, at a higher level, the CoE networks that we constructed have captured most topological properties and functional relationships in the true network. We expect that a more sophisticated method for constructing CoE networks, such as Bayesian networks [41] and Boolean networks [42], may improve the discovery of function modules even further.

The CoE networks that we constructed posses a unique topological feature that is different from the CoE networks reported in the literature. In our network, the exponent of the power-law degree distribution falls in the range of 2 to 3, similar to most other real-world networks, whereas the exponent of CoE networks reported in the literature is below the critical value of 2. We are currently looking for the causes of this discrepancy and examining their effects on our clustering algorithm.

Although we have only demonstrated our method on gene expression data, it can be applied to other types of experimental data as well. The efficiency of our clustering method and its relative independence of any detailed domain knowledge of the data make it well suited for identifying intrinsic structures in large-scale network data. Furthermore, the cluster evaluation method we proposed may be used as a general framework for assessing different algorithms and comparing clustering results based on external knowledge.

Acknowledgements

This research was supported in part by an NSF grants ITR/EIA-0113618 and IIS-0535257, and a grant from Monsanto Company.

References

1. Tong, A., Drees, B., Nardelli, G., Bader, G., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C., Fields, S., Boone, C., Cesareni, G.: A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295** (2002) 321–324
2. Stuart, J., Segal, E., Koller, D., Kim, S.: A gene-coexpression network for global discovery of conserved genetic modules. Science **302** (2003) 249–255
3. Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabasi, A.: The large-scale organization of metabolic networks. Nature **407** (2000) 651–654

4. Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., Young, R.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science **298** (2002) 799–804
5. Jeong, H., Mason, S., Barabasi, A., Oltvai, Z.: Lethality and centrality in protein networks. Nature **411** (2001) 41–42
6. Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabasi, A.: Hierarchical organization of modularity in metabolic networks. Science **297** (2002) 1551–1555
7. Barabasi, A., Oltvai, Z.: Network biology: understanding the cell's functional organization. Nat Rev Genet **5** (2004) 101–113
8. Oltvai, Z., Barabasi, A.: Systems biology. life's complexity pyramid. Science **298** (2002) 763–764
9. Armstrong, N., van de Wiel, M.: Microarray data analysis: from hypotheses to conclusions using gene expression data. Cell Oncol. **26** (2004) 279–290
10. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95** (1998) 14863–14868
11. Tavazoie, S., Hughes, J., Campbell, M., Cho, R., Church, G.: Systematic determination of genetic network architecture. Nat. Genet. **22** (1999) 281–285
12. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T.: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA **96** (1999) 2907–2912
13. Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., Gerstein, M.: Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. J. Mol. Bio. **314** (2001) 1053–1066
14. Bolshakova, N., Azuaje, F.: Machaon CVE: cluster validation for gene expression data. Bioinformatics **19** (2003) 2494–2495
15. Azuaje, F., Al-Shahrour, F., Dopazo, J.: Ontology-driven approaches to analyzing data in functional genomics. Methods Mol Biol. **316** (2006) 67–86
16. Gibbons, F., Roth, F.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. **12** (2002) 1574–1581.
17. Ruan, J., Zhang, W.: Identification and evaluation of weak community structures in networks. In: Proc. National Conf. on AI, (AAAI-06). (2006) 470–475
18. Ruan, J., Zhang, W.: Discovering weak community structures in large biological networks. Technical Report cse-2006-20, Washington University in St Louis (2006)
19. Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. Nucleic Acids Res. **34** (2006) D535–539
20. The Gene Ontology Consortium: The gene ontology (GO) database and informatics resource. Nucleic Acids Res **32** (2004)
21. Harbison, C., Gordon, D., Lee, T., Rinaldi, N., Macisaac, K., Danford, T., Hannett, N., Tagne, J., Reynolds, D., Yoo, J., Jennings, E., Zeitlinger, J., Pokholok, D., Kellis, M., Rolfe, P., Takusagawa, K., Lander, E., Gifford, D., Fraenkel, E., Young, R.: Transcriptional regulatory code of a eukaryotic genome. Nature. **431** (2004) 99–104
22. Zhou, X., Kao, M., Wong, W.: Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci U S A **99** (2002) 12783–12788
23. Carter, S., Brechbuhler, C., Griffin, M., Bond, A.: Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics **20** (2004) 2242–2250

24. Zhu, D., Hero, A., Cheng, H., Khanna, R., Swaroop, A.: Network constrained clustering for gene microarray data. Bioinformatics **21** (2005) 4014–4020
25. Aggarwal, A., Guo, D., Hoshida, Y., Yuen, S., Chu, K., So, S., Boussioutas, A., Chen, X., Bowtell, D., Aburatani, H., Leung, S., Tan, P.: Topological and functional discovery in a gene coexpression meta-network of gastric cancer. Cancer Res **66** (2006) 232–241
26. Fjallstrom, P.: Algorithms for graph partitioning: A survey. Linkoping Electron. Atricles in Comput. and Inform. Sci. (1998)
27. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys **69** (2004) 026113
28. Newman, M.: The structure and function of complex networks. SIAM Review **45** (2003) 167–256
29. Danon, L., Duch, J., Diaz-Guilera, A., Arenas, A.: Comparing community structure identification. J. Stat. Mech. (2005) P09008
30. Garey, M., Johnson, D.: Computers and Intractability: A Guide to the Theory of NP-completeness. Freeman, San Francisco (1979)
31. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS. (2001) 849–856
32. White, S., Smyth, P.: A spectral clustering approach to finding communities in graph. In: SIAM Data Mining. (2005)
33. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 888–905
34. Altman, D.: Practical Statistics for Medical Research. Chapman & Hall/CRC (1991)
35. Boyle, E., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J., Sherlock, G.: Go::termfinder - open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics **20** (2004) 3710–3715
36. Jones, K.S.: Idf term weighting and ir research lessons. Journal of Documentation **60** (2004) 521–523
37. Gasch, A., Spellman, P., Kao, C., Carmel-Harel, O., Eisen, M., Storz, G., Botstein, D., Brown, P.: Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell **11** (2000) 4241–4257
38. Albert, R., Barabasi, A.: Statistical mechanics of complex networks. Reviews of Modern Physics **74** (2002) 47
39. Saccharomyces genome database. `http://www.yeastgenome.org/`
40. Wallace, D.L.: Comment. Journal of the American Statistical Assocation **78** (1983) 569–576
41. Friedman, N., Linial, M., Nachman, I., Peer, D.: Using bayesian networks to analyze expression data. J Comput Biol. **7** (2000) 601–620
42. Kauffman, S.: A proposal for using the ensemble approach to understand genetic regulatory networks. J Theor Biol. **230** (2004) 581–590