Report Number: WUCSE-2006-31

2006-01-01

Discovering Functional Modules by Clustering Gene Co-expression Networks

Jianhua Ruan and Weixiong Zhang

Identification of groups of functionally related genes from high throughput gene expression data is an important step towards elucidating gene functions at a global scale. Most existing approaches treat gene expression data as points in a metric space, and apply conventional clustering algorithms to identify sets of genes that are close to each other in the metric space. However, they usually ignore the topology of the underlying biological networks. In this paper, we propose a network-based clustering method that is biologically more realistic. Given a gene expression data set, we apply a rank-based transformation to obtain a sparse co-expression...
Read complete abstract on page 2.

Recommended Citation

Ruan, Jianhua and Zhang, Weixiong, "Discovering Functional Modules by Clustering Gene Co-expression Networks" Report Number: WUCSE-2006-31 (2006). *All Computer Science and Engineering Research.*
https://openscholarship.wustl.edu/cse_research/182

Department of Computer Science & Engineering - Washington University in St. Louis
Campus Box 1045 - St. Louis, MO - 63130 - ph: (314) 935-6160.

Discovering Functional Modules by Clustering Gene Co-expression Networks

Jianhua Ruan and Weixiong Zhang

Complete Abstract:

Identification of groups of functionally related genes from high throughput gene expression data is an important step towards elucidating gene functions at a global scale. Most existing approaches treat gene expression data as points in a metric space, and apply conventional clustering algorithms to identify sets of genes that are close to each other in the metric space. However, they usually ignore the topology of the underlying biological networks. In this paper, we propose a network-based clustering method that is biologically more realistic. Given a gene expression data set, we apply a rank-based transformation to obtain a sparse co-expression network, and use a novel spectral clustering algorithm to identify natural community structures in the network, which correspond to gene functional modules. We have tested the method on two large-scale gene expression data sets in yeast and Arabidopsis, respectively. The results show that the clusters identified by our method on these datasets are functionally richer and more coherent than the clusters from the standard k-means clustering algorithm.

2006-31

Discovering Functional Modules by Clustering Gene Co-expression Networks

Authors: Jianhua Ruan, Weixiong Zhang

Corresponding Author: jruan@cse.wustl.edu

Abstract: Identification of groups of functionally related genes from high throughput gene expression data is an important step towards elucidating gene functions at a global scale. Most existing approaches treat gene expression data as points in a metric space, and apply conventional clustering algorithms to identify sets of genes that are close to each other in the metric space. However, they usually ignore the topology of the underlying biological networks. In this paper, we propose a network-based clustering method that is biologically more realistic. Given a gene expression data set, we apply a rank-based transformation to obtain a sparse co-expression network, and use a novel spectral clustering algorithm to identify natural community structures in the network, which correspond to gene functional modules. We have tested the method on two large-scale gene expression data sets in yeast and Arabidopsis, respectively. The results show that the clusters identified by our method on these datasets are functionally richer and more coherent than the clusters from the standard k-means clustering algorithm.

Type of Report: Other

Discovering Functional Modules by Clustering Gene Co-expression Networks

Jianhua Ruan[1] and Weixiong Zhang[1,2]
Department of Computer Science[1] and Department of Genetics[2]
Washington University in St. Louis, St. Louis, MO 63130, USA
jruan@cse.wustl.edu, zhang@cse.wustl.edu

Abstract

Identification of groups of functionally related genes from high throughput gene expression data is an important step towards elucidating gene functions at a global scale. Most existing approaches treat gene expression data as points in a metric space, and apply conventional clustering algorithms to identify sets of genes that are close to each other in the metric space. However, they usually ignore the topology of the underlying biological networks. In this paper, we propose a network-based clustering method that is biologically more realistic. Given a gene expression data set, we apply a rank-based transformation to obtain a sparse co-expression network, and use a novel spectral clustering algorithm to identify natural community structures in the network, which correspond to gene functional modules. We have tested the method on two large-scale gene expression data sets in yeast and Arabidopsis, respectively. The results show that the clusters identified by our method on these datasets are functionally richer and more coherent than the clusters from the standard k-means clustering algorithm.

Keywords: clustering, Microarray, co-expression network

1 Introduction

Many biological sub-systems considered in systems biology can be modeled as networks, where nodes are such entities as genes or proteins, and edges are the relationships between pairs of entities. Examples of biological networks include protein-protein interaction networks [23], gene co-expression networks [20], metabolic networks [11], and transcriptional regulatory networks [13]. Much effort has been devoted to the study of their overall topological properties and similarities to other real-world networks [10, 19, 4, 17].

A large amount of available gene expression microarray data has provided opportunities for identifying gene functions on a global scale. Since genes that are on the same pathways or form functional complexes are often co-regulated, they often exhibit similar expression patterns under diverse temporal and physiological conditions. Therefore, genes are often clustered according to their expression patterns for gene function analysis. The most popular clustering techniques include hierarchical clustering [6], k-means clustering [22], and self-organizing maps [21].

However, genes of similar expression patterns may not necessarily have similar functions. Genes could be accidentally co-regulated [20]; a single event often activate multiple pathways that have distinct biological functions. On the other hand, genes with related functions may not share close correlation in their expression patterns. For example, there might be time-shift between the expression patterns of genes in the same pathway [18].

Here, we take a network-based perspective toward gene expression clustering. Given gene expression data, We first construct a co-expression network, where the nodes in the network are genes, and the edges reflect co-expression relationships between pairs of genes. We then develop a network clustering algorithm to identify subsets of genes that are relatively densely connected to one another. Using gene expression datasets on yeast and Arabidopsis under various stress conditions, we show that the clusters of genes obtained from our method are functionally richer and more coherent than that obtained from the gene expression data directly.

The paper is organized as follows. In section 2, we described our methods for constructing gene co-expression networks and the network clustering algorithm. In section 3, we present our cluster results and compare them with the results from the standard k-means algorithm. We conclude in section 4.

2 Methods

Our method for identifying functional modules from gene expression data consists of three main steps. First, we construct a co-expression network from the expression data. Second, we apply an algorithm that we have recently developed to cluster the co-expression network into densely connected sub-graphs. The algorithm was designed specifically for clustering networks, and is able to automatically determine the most appropriate number of clusters. In the final step, we analyze the enriched functional categories for genes in each cluster, and assign putative functions to unknown genes according to the cluster they belong to.

2.1 Constructing gene co-expression networks

Several methods have been recently proposed for constructing a co-expression network from gene expression data. The most straightforward method first calculates some similarity measure between the expression profiles of every pair of genes, and determines a cut-off value to select pairs of genes that should be connected [24]. The problem with this approach, aside being arbitrary in choosing the cut-off, is that gene expression correlation coefficient values often exhibit some local-scaling properties. That is, some sets of genes are correlated to each other with very high correlation coefficients, while some other genes are only loosely connected to each other via medium or low correlation coefficients. Therefore, if we choose a cut-off too stringent, the genes in the latter set will become disconnected from the network. On the other hand, if we let every gene be connected to the network, the cut-off might be so low that a large fraction of genes are almost completely connected, making further partitioning a difficult task.

Here, we propose a rank-based transformation of similarity matrices to deal with such local-scaling property. To this end, we calculate Pearson correlation coefficients between every pair of genes, and for every gene, rank all other genes by its correlation coefficient to the former. 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, $R(j, i)$. We then decide a threshold α to select co-expression links that are ranked within top-α with respect to some gene. By varying α, we can obtain networks of different granularity. A network constructed as this is directed, but we will ignore the directions when clustering the network.

The above rank-based construction may seem too simple at a first glance. Note that, however, our purpose here is not to construct a network of some optimality. That problem is often attacked by algorithms such as Bayesian networks and Boolean networks [7, 12], which seek for networks that are optimal given the data. Here, our main focus is to construct a representative network that would facilitate our clustering algorithm to discover functional modules. Later in the Results section we will show that clustering on such networks is indeed biologically more meaningful than clustering the complete similarity matrix as in traditional clustering. We will also show that clustering on this seemingly arbitrary network is rather robust, in that perturbing a large fraction of its connections does not affect the final clustering results significantly.

The idea of using rank-transformation to construct co-expression networks has been adopted previously by [20, 1]. In their studies, they applied statistical analysis to choose co-expression links whose ranks were consistently high in multiple data sources or multiple organisms. Given only a single data set, we have tried to use bootstrapping to remove some of the high-ranked edges that may be due to noises. We found that, in general, such statistical treatment did improve clustering quality to a limited extent. As these are not the focus of this paper, we will ignore them in the subsequent discussion.

2.2 Network clustering

We have recently proposed a spectral clustering algorithm that was designed specifically for networks. The method has several unique features. First, it considers local neighborhood information for any node, and therefore greatly improves clustering quality of networks. Second, it combines a modularity function Q to automatically determine the most appropriate number of clusters in a network, which is a difficult problem for any clustering algorithm. Third, a greedy algorithm has been developed to recursively partition a network to optimize Q. The greedy algorithm can handle networks of several thousands of nodes in a few minutes, several orders of magnitude faster than a previous algorithm [15], while often achieving comparable clustering qualities. We have tested the algorithm extensively on many simulated networks and networks with known structures, as well as several real applications such as protein-protein interaction networks and scientific collaboration networks, all indicating that our method is both efficient and effective.

The detailed description and analysis of the algorithm will be reported elsewhere. Here we briefly describe the key steps in the algorithm. Given a graph or network $G = (V, E)$, where V is a set of nodes, and E a set of edges, let $A = (A_{ij})$ be the adjacency matrix of G, i.e. $A_{ij} = 1$ if $(v_i, v_j) \in E$, or 0 otherwise. Let D be the diagonal degree matrix of A, where $D_{ii} = \sum_j A_{ij}$. Further define

$$\begin{aligned} B &= D^{-1/2} \times A^2 \times D^{-1/2} \cdot (1 - I), \\ C &= D^{-1/2} \times A^2 \times D^{-1/2} \cdot A \cdot (1 - I), \text{ and} \\ H &= \alpha \times A + \beta \times B + C, \end{aligned}$$

where I is an identity matrix, "$\times$" represents ordinary matrix multiplication and "$\cdot$" means entry-wise multiplication.

The matrices B and C compute the numbers of length-2 paths and triangles connecting every pair of nodes, respectively, scaled by the number of edges that each node has. Therefore, they captures some local neighborhood information of each node. Node pairs within the same cluster often have higher weights in the B or C matrix than those belonging to different clusters. The matrix H is a combination of A, B and C, while α and β are two free variables. We have found that in most cases, clustering the H matrix by taking the values of α and β such that $\alpha \times A$, $\beta \times B$ and C have the same maximal weight can results in the best clustering accuracy.

To determine the most appropriate number of clusters, we adopt a modularity function, Q, proposed by Newman and Girvan [15], which is defined as:

$$Q(\Gamma_k) = \sum_{i=1}^{k} (e_{ii} - a_i^2),$$

where Γ_k is a particular clustering that partitions a graph into k groups, e_{ii} is the fraction of the edges that fall within cluster i, and a_i the fraction of edges each of which has at least one end connecting to cluster i. Intuitively, the Q function measures the percentage of edges fully contained within clusters, subtracted by the percentage that one would expect by chance. Empirically, Newman and Girvan have shown that higher Q values correspond to better clusters in general.

The spectral greedy algorithm *k-cuts* works as follows. Given the derived matrix H from a network G, we apply the standard spectral clustering algorithm [16] to obtain $k = 2, 3, \cdots, K$ clusters, where K is a small integer ($K < 10$ typically). For each clustering Γ_k returned, we calculate the Q measurement, and pick the Γ_k that gives the best Q value. Then, for each cluster that has not be tried, we recursively partition it into $k = 2, 3, \cdots, K$ clusters, and measure the Q value on the full network. If the Q value is improved, we accept the partition; otherwise we move on to the next cluster.

2.3 Functional analysis

To assess the functional significance of obtained gene clusters, and suggest putative functions for genes with unknown functions, we calculate the enrichment of gene ontology (GO) [9] molecular function, biological process and cellular component terms for the genes within each cluster. The significance of enrichment is measured by an accumulative hypergeometric test [3], and the P-values are adjusted by Bonferroni corrections for multiple tests [3]. The search of the GO trees is performed with a computer program GO:TermFinder [5].

3 Results

3.1 Clustering the yeast co-expression network

We first applied our method to cluster a co-expression network in the budding yeast. We obtained the yeast gene expression data measured in 173 different time points under various stress conditions [8]. We selected 3000 genes that showed the most variations in their expression data, and constructed a network with $\alpha = 2$, 3, 4 and 5, respectively. That is, we let each gene connect to its top-α correlated genes (see section 2.1). Some statistics about the networks are listed in Table 2. We then applied the *k-cuts* algorithm to cluster each network. The best numbers of clusters for the four networks are 24, 20, 12 and 12, respectively.

To validate the biological significance of the clusters we had, we counted the number of GO terms enriched in the clusters at various significance levels. To rule out the possibility that a single cluster may contain a very large number of enriched GO terms and therefore predominates the contribution from other clusters, we also counted the number of clusters that had at least one enriched GO term at a given significance level. For comparison, we also applied the standard k-means algorithm directly on the expression data, using Pearson correlation-coefficient as the distance measure, and specified the number of clusters k = 24, 20, 12 and 12 respectively. Furthermore, we randomly shuffled our clustering results for the $\alpha = 3$ network by fixing the size of each cluster and randomly assigning genes to clusters.

Fig. 1 shows the results of the GO analysis on these clustering results. As can be seen, comparing to random clustering, our method and the k-means algorithm both identified strongly functionally correlated clusters. More importantly, the gene clusters identified by our methods contain significantly higher number of GO terms than the standard k-means algorithm at all significance levels and for different number of clusters (Fig. 1(a)-(d)). Furthermore, the numbers of clusters containing at least one enriched GO terms are also larger for our methods than for k-means (Fig. 1(e)-(h)).

Table 1 shows the number of genes within each

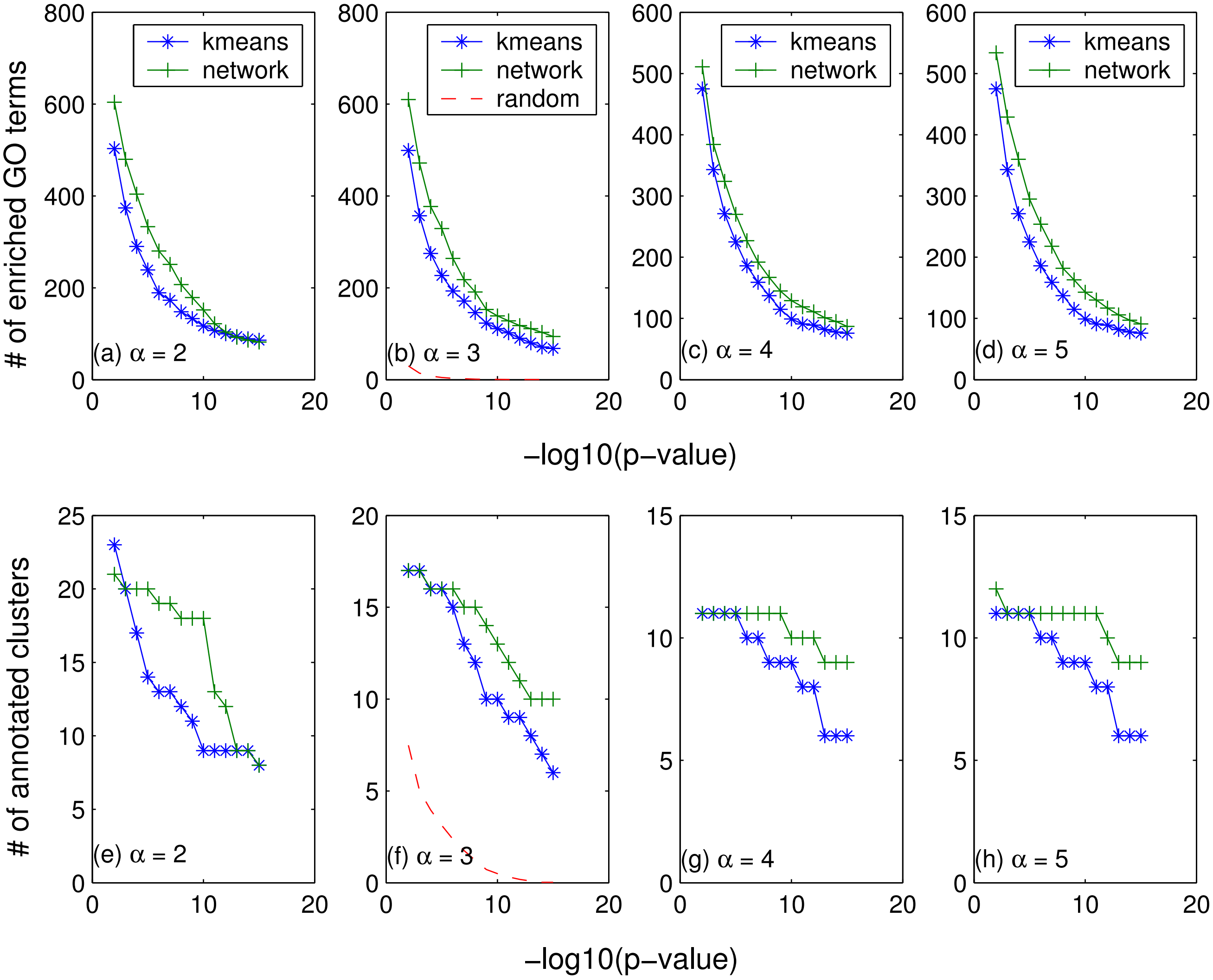

Figure 1. Enrichment of GO terms in yeast co-expression network clusters. The legends in (e)-(h) are the same as in (a)-(d), respectively.

cluster and the most significant GO terms associated with each cluster. As can been seen, most clusters contain highly coherent functional groups, e.g. clusters 1 (retrotransposon nucleocapsid), 6 (ribosome), 7 (ribosome biogenesis) and 13 (generation of precursor metabolites and energy). Our algorithm identified several small clusters with size < 15. Interestingly, those small clusters correspond to very specific functional groups. For example, 8 of 10 genes in cluster 2 are in nuclear nucleosome, while there are a total of 12 in the genome, an enrichment of 480 folds; cluster 11 contains 4 of the 7 galactose metabolism genes, an enrichment of 823 folds. More than half of the genes in two small clusters (8 and 14) have unknown cellular components and no other enriched GO terms. It is very likely that those two clusters represent specific functional modules that have not be studied. Several large clusters (3, 15 and 16) contain both a large fraction of genes with unknown functions, and groups of genes with significantly enriched common functions. Cluster 3 contains 42 genes with oxidoreductase activities and 53 gene responsive to stress, while the functions of many other genes in the cluster are unknown. It is possible that these genes also have similar functions.

Since gene expression measurement contains some inherent variability, and our method only used the top-ranked co-expression links to construct the co-expression network, we wanted to evaluate whether the clusters were stable with respect to perturbations in the network structure. To evaluate this, we removed all the top-ranked co-expression links from the $\alpha = 3$ network. That is, each gene is now connected only to its second and third-best correlated genes. This network has about the same connectivity as the $\alpha = 2$ network, but very different connections. Surprisingly, most of the clusters are very similar to those obtained from the $\alpha = 2$ network, and 55% of the gene-pairs are conserved between the two clusters. Furthermore, the clusters still contain significantly more enriched GO terms than the clusters identified by k-means (data not shown).

Previous studies have analyzed the topologies of various real networks, including biological networks, and suggested a common scale-free property [10, 19, 4, 17]. In a scale-free network, the probability for a node to have n edges obeys the power-law distribution, i.e. $P(n) = c \times n^{-\gamma}$, where c is a constant. The result of the scale-free property is that a few nodes in the network are highly connected, acting as hubs, while most nodes are of low degree. In contrast, in a random network where the connections are spread uniformly most nodes have similar degrees. Real networks also differ in random network in that they often have high clustering coefficient [14].

To determine the topological characters of the co-expression network, we plotted the number of genes having n connections as a function of n in a log-log scale. To compare, we constructed networks based on a randomly shuffled version of the original gene expression data. As shown in Fig. 2, the real co-expression networks exhibit 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 co-expression network. In comparison, the co-expression networks constructed from the randomized expression data are more similar to random networks and contain much smaller number of high-degree nodes.

Second, we calculated the clustering coefficients of the co-expression networks derived from true expression data and randomly shuffled expression data. As shown in Table 2, the true co-expression networks have much higher clustering coefficients than the random network, indicating that the co-expression networks are highly modular. To ensure that the high clustering coefficient is not an artifact of scale-free networks, we permuted the co-expression networks through random rewiring [2]. The rewiring procedure preserves the degree for each node, thus does not change the scale-free property of the networks. As shown in Table 2, the clustering coefficients of the rewired networks are significantly lower than

Table 1. Functional modules in the yeast co-expression network.

cluster	size	GO term	Genes in cluster/genome	Enrichment	P-value*
1	21	retrotransposon nucleocapsid	17 / 94	62.0	1.2E-29
2	10	nuclear nucleosome	8 / 12	480.0	1.2E-22
3	514	biological process unknown	200 / 1772	1.6	6.3E-14
		molecular function unknown	234 / 2393	1.4	1.3E-09
		oxidoreductase activity	42 / 235	2.5	1.9E-08
		response to stress	53 / 350	2.1	8.9E-08
4	144	telomere maintenance	6 / 35	8.6	5.8E-05
5	12	asparagine catabolism	4 / 5	480.0	2.2E-11
6	206	ribosome	124 / 276	15.7	4.0E-132
7	553	ribosome biogenesis	127 / 196	8.4	1.7E-96
8	17	cellular component unknown	9 / 1063	3.6	2.6E-04
9	63	amino acid metabolism	21 / 176	13.6	4.8E-19
10	11	helicase activity	9 / 83	71.0	1.2E-16
11	5	galactose metabolism	4 / 7	822.9	1.6E-12
12	205	macromolecule catabolism	34 / 231	5.2	9.5E-16
13	140	generation of precursor metabolites and energy	46 / 216	11.0	1.4E-36
14	15	cellular component unknown	8 / 1063	3.6	5.4E-04
15	216	molecular function unknown	106 / 2393	1.5	7.4E-07
		monosaccharide metabolism	13 / 89	4.9	2.1E-06
16	298	cellular component unknown	112 / 1063	2.5	2.4E-23
		molecular function unknown	159 / 2393	1.6	2.3E-13
		biological process unknown	126 / 1772	1.7	6.1E-12
		spore wall assembly	9 / 24	9.1	2.4E-07
		vitamin metabolism	14 / 69	4.9	6.3E-07
		pyridoxine metabolism	5 / 7	17.3	2.3E-06
17	97	nitrogen compound metabolism	18 / 127	5.9	8.3E-10
18	23	aryl-alcohol dehydrogenase activity	5 / 8	195.7	1.2E-11
19	435	catalytic activity	186 / 1853	1.7	1.2E-15
		cellular localization	69 / 464	2.5	7.5E-13
20	15	purine base metabolism	6 / 15	192.0	1.3E-13

*the P-values shown here were not adjusted the original networks, indicating that high modularity is indeed a property of the co-expression networks.

It is not surprising to find out that co-expression network is yet another example of scale-free networks. However, several previous studies on a number of gene co-expression networks have suggested that there might exist formal topological differences between gene co-expression networks and other biological networks [20, 1]. In these studies, it has been observed that the exponent γ for the power law degree distribution of co-expression networks was consistently less than 2, while in most other scale-free networks, γ is within the range $[2, 3]$ (see the networks listed in [14, 2] for examples). As have been proved [14], scale-free networks with $\gamma < 2$ have no finite mean degree when the network grows to infinity, and is dominated by nodes with large degrees. To determine the γ values for the co-expression networks we have constructed, we fitted a linear regression to each log-log plot to calculate the slope. The γ values for different networks are shown together with the fitted lines in Fig. 2. As can be seen, γ is consistently within the range [2, 3] in our networks, similar to most other real-world networks and biological networks such as protein interaction networks and metabolic networks. This apparent similarity to other types of real networks but difference to previous co-expression networks may have been caused by our method to select the co-expression links. Although more work is required, we speculate that the networks constructed by our methods may better represent the underlying biological networks than previous co-expression networks.

3.2 clustering the plant cold-stress regulated genes

To see if our network clustering method also works for higher organisms, we applied it to a co-expression network of Arabidopsis genes. We downloaded the normalized expression data of Arabidopsis genes from the AtGenExpress database (`http://www.uni-tuebingen.de/plantphys/AFGN/atgenex.htm`). The data set contains the expression data of $\approx$ 22k Arabidopsis genes in the root or shoot tissues in 12 time points following cold stress treatment. We selected the genes that are up- or down-regulated by at least four folds in at least one of the 12 time points. We constructed the co-expression network by connecting each gene to its top three correlated genes ($\alpha = 3$). We then made the network undirected by ignoring the directions. This process produces a network with 2545 genes and 5838 co-expression links.

Our clustering algorithm partitioned the network into 19 clusters, with a Q value 0.81, indicating strong modular structures. We counted the number of GO terms enriched in the clusters at various significance levels, as in the previous experiments. We also applied the standard k-means algorithm to cluster the gene expression data into 19 clusters and repeated the GO analysis.

Fig. 3 shows the number of enriched GO terms in the clusters with respect to the genes in the network, and the total number of clusters with at least one enriched GO term at various significance levels. As can be seen, the clusters identified by our network-based clustering algorithm are functionally more coherent than that identified by the k-means algorithm, similar to what we observed in the yeast co-expression network. Table 3 shows the most enriched functional categories for each cluster. Some clusters 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).

4 Conclusions and discussion

In this paper, we proposed a network-based method for clustering microarray gene expression data. We introduced a simple rank-based method to construct gene co-expression networks from

Table 2. Summary of network statistics.

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
m	5432	8103	10775	13432
k_{avg}	1.8	2.7	3.6	4.5
c	0.089	0.124	0.144	0.159
c_{random}	$0.010 \pm .002$	$0.015 \pm .002$	$0.018 \pm .001$	$0.020 \pm .001$
$c_{scalefree}$	$0.002 \pm .0002$	$0.003 \pm .0001$	$0.004 \pm .0001$	$0.005 \pm .0001$

m: number of edges; k_{avg}: averge node degree; c: clustering coefficient; c_{random}: clustering coefficient of the network constructed from permuted expression data; $c_{scalefree}$: clustering coefficient of the rewired network.

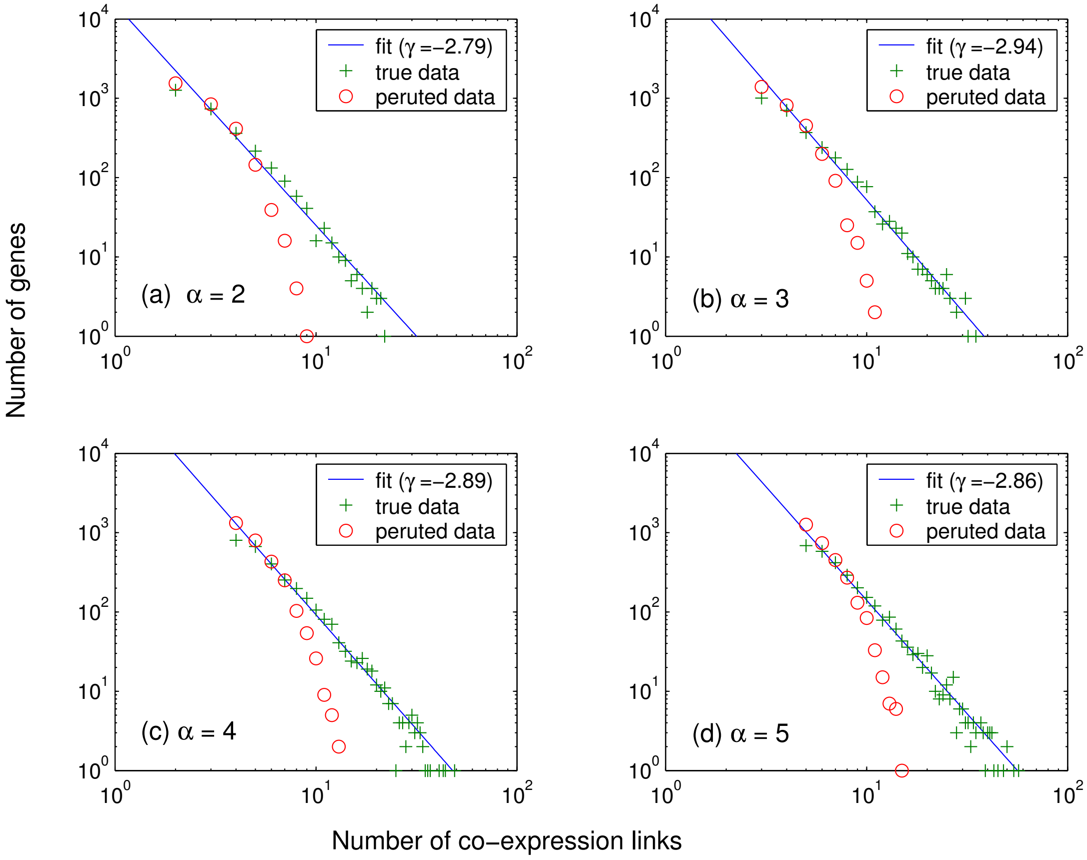

Figure 2. Distribution of the number of co-expression links for each gene.

Table 3. Functional modules in the Arabidopsis co-expression network.

cluster	size	GO term	Genes in cluster/network	Enrichment	P-value*
1	199	-	-	-	-
2	141	-	-	-	-
3	79	-	-	-	-
4	180	catalytic activity	99 / 1133	1.6	4.1E-09
		amino acid and derivative metabolism	18 / 74	4.4	3.9E-08
5	284	endomembrane system	79 / 572	1.6	3.5E-06
6	238	oxidoreductase activity	40 / 214	2.6	7.7E-09
		secondary metabolism	18 / 80	3.1	9.9E-06
7	65	photosynthesis	11 / 17	32.6	8.7E-16
8	261	RNA binding	11 / 30	4.6	9.2E-06
9	186	galactolipid biosynthesis	3 / 3	17.6	1.8E-04
10	19	branched-chain-amino-acid transaminase activity	3 / 3	172.6	1.7E-07
11	117	starch metabolism	4 / 7	16.0	5.0E-05
		circadian rhythm	6 / 22	7.6	8.5E-05
12	271	protein modification	37 / 210	2.1	4.3E-06
13	268	methyltransferase activity	8 / 21	4.7	1.4E-04
14	13	response to heat	8 / 23	87.7	1.9E-15
15	223	antiporter activity	10 / 24	6.1	1.5E-06
16	151	transcription regulator activity	60 / 428	3.0	2.5E-17
17	200	zeaxanthin epoxidase activity	3 / 3	16.4	2.2E-04
18	17	lipid binding	5 / 20	48.2	2.9E-08
		membrane	12 / 869	2.7	1.8E-04
19	249	calcium ion binding	13 / 53	3.2	1.1E-04

*the P-values shown here were not adjusted

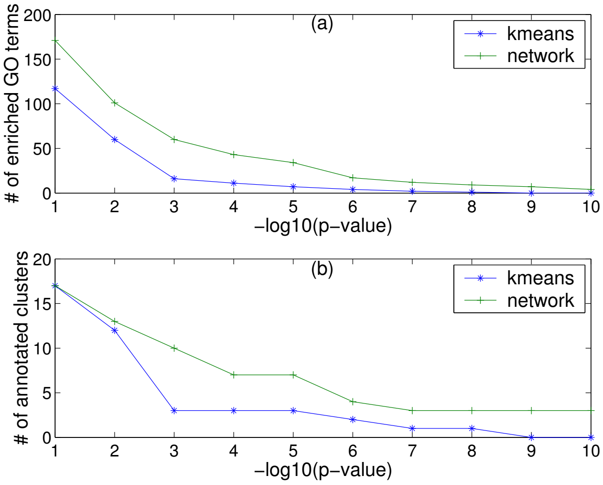

Figure 3. Enrichment of GO terms in the Arabidopsis co-expression network clusters.

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 co-expression networks in yeast and Arabidopsis, respectively, and showed the new network-based clustering can produce biologically more meaningful clusters than traditional methods such as k-means. The clusters identified by our methods always contain more significantly enriched GO terms than the k-means algorithm.

It is rather surprising that the simple method we proposed to construct co-expression networks worked well. The connections in such a co-expression network are obviously very different from exact biological interactions. Nevertheless, at a higher level, the co-expression network we constructed can capture most topological properties in the true underlying network. Genes on the same pathway tend to be close to one another in the co-expression network and vice versa. We expect that a more sophisticated method for constructing co-expression networks will improve the discovery of function modules even further.

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

Acknowledgements

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

References

[1] A. Aggarwal, D. Guo, Y. Hoshida, S. Yuen, K. Chu, S. So, A. Boussioutas, X. Chen, D. Bowtell, H. Aburatani, S. Leung, and P. Tan. Topological and functional discovery in a gene coexpression meta-network of gastric cancer. *Cancer Res*, 66(1):232–41, Jan 2006.

[2] R. Albert and A. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47, 2002.

[3] D. G. Altman. *Practical Statistics for Medical Research*. Chapman & Hall/CRC, 1991.

[4] A. Barabasi and Z. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–13, Feb 2004.

[5] E. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. Cherry, and G. Sherlock. Go::termfinder - open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, pages D258–61, Aug 2004.

[6] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–8, 1998.

[7] N. Friedman, M. Linial, I. Nachman, and D. Peer. Using bayesian networks to analyze expression data. *J Comput Biol.*, 7:601–20, 2000.

[8] A. Gasch, P. Spellman, C. Kao, O. Carmel-Harel, M. Eisen, G. Storz, D. Botstein, and P. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, Dec 2000.

[9] M. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. Rubin, J. Blake, C. Bult, M. Dolan, H. Drabkin, J. Eppig, D. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. Cherry, K. Christie, M. Costanzo, S. Dwight, S. Engel, D. Fisk, J. Hirschman, E. Hong, R. Nash, A. Sethuraman, C. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The gene ontology (go) database and informatics resource. *Nucleic Acids Res*, 32 Database issue, Jan 2004.

[10] H. Jeong, S. Mason, A. Barabasi, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–2, May 2001.

[11] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–4, Oct 2000.

[12] S. Kauffman. A proposal for using the ensemble approach to understand genetic regulatory networks. *J Theor Biol.*, 230:581–90, 2004.

[13] T. Lee, N. Rinaldi, F. Robert, D. Odom, Z. Bar-Joseph, G. Gerber, N. Hannett, C. Harbison, C. Thompson, I. Simon, J. Zeitlinger, E. Jennings, H. Murray, D. Gordon, B. Ren, J. Wyrick, J. Tagne, T. Volkert, E. Fraenkel, D. Gifford, and R. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct 2002.

[14] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[15] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2):026113, Feb 2004.

[16] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[17] Z. Oltvai and A. Barabasi. Systems biology. life's complexity pyramid. *Science*, 298(5594):763–4, Oct 2002.

[18] J. Qian, M. Dolled-Filhart, J. Lin, H. Yu, and M. Gerstein. Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Bio.*, 314:1053–66, Dec 2001.

[19] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, Aug 2002.

[20] J. Stuart, E. Segal, D. Koller, and S. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–55, Oct 2003.

[21] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–12, 1999.

[22] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22:281–5, 1999.

[23] A. Tong, B. Drees, G. Nardelli, G. Bader, B. Brannetti, L. Castagnoli, M. Evangelista, S. Ferracuti, B. Nelson, S. Paoluzi, M. Quondam, A. Zucconi, C. Hogue, S. Fields, C. Boone, and G. Cesareni. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, 295(5553):321–4, Jan 2002.

[24] X. Zhou, M. Kao, and W. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad Sci U S A*, 99:12783–8, 2002.