Available online at www.sciencedirect.com
ScienceDirect
BioSystems 91 (2008) 158–165
www.elsevier.com/locate/biosystems

Using matrix of thresholding partial correlation coefficients to infer regulatory network☆

Lide Han[a,b], Jun Zhu[a,*]

[a] Institute of Bioinformatics, College of Agriculture & Biotechnology, Zhejiang University, Hangzhou, Zhejiang 310029, China
[b] College of Agronomy, Anhui Agricultural University, Hefei, Anhui 230036, China

Received 7 March 2007; received in revised form 24 August 2007; accepted 24 August 2007

Abstract

DNA arrays measure the expression levels for thousands of genes simultaneously under different conditions. These measurements reflect many aspects of the underlying biological processes. A method based on the matrix of thresholding partial correlation coefficients (MTPCC) is proposed for network inference from expression profiles. It includes three main parts: (1) hierarchical cluster analysis, (2) cluster boundaries establishment, and (3) regulatory network inference. The method was applied to the expression data of 2467 genes in *Saccharomyces cerevisiae* measured under 79 different conditions [Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95, 14863–14868]. Using hierarchical clustering and cluster boundaries establishment, the 2467 genes were grouped into 12 clusters. The expression profiles of each cluster were expressed as a set of expression levels average over the cluster that constituted genes of each condition. Then the expression data of these clusters were subjected to the analysis of partial correlation, and the significance of each element in the obtained partial correlation coefficient matrix (PCCM) was examined by a permutation test. The corresponding undirected dependency graph (UDG) was obtained as a model of the regulatory network of *S. cerevisiae*. The veracity of the network was evidenced by the consistency of our results with the collected results from experimental studies.

Keywords: Network inference; Partial correlation coefficient; Permutation test; UDG; Microarray

1. Introduction

A DNA microarray, either an oligonucleotide array (e.g. Affymetrix) or a cDNA array, can be used to measure relative expression levels of thousands of genes simultaneously in biological samples (cells, tissues, tumors, etc.) under various conditions (DeRisi et al., 1997). Usually, such experiments are designed to reflect many aspects of biological processes of interest (Spellman et al., 1998). However, the sheer amount of data presents a challenge in developing effective methods that are both statistically sound and computationally tractable, in particular for inferring biological interactions.

Various methods, for example, Boolean Networks (Somogyi and Shiegoski, 1996; Akutsu et al., 1999), Bayesian Networks (Friedman et al., 2000; Hartemink et al., 2002) and Dynamic Bayesian Networks (Murphy and Mian, 1999), were proposed to infer the regulatory network from expression profiles. These methods

☆ Availability: a program Network_Inference 1.0 written in C^{++} is available by contacting jzhu@zju.edu.cn.

* Corresponding author. Tel.: +86 571 86971731; fax: +86 571 86971498.
E-mail address: jzhu@zju.edu.cn (J. Zhu).

0303-2647/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2007.08.008

have several limitations, including the discretization for the gene expression levels leading to loss of information, the need for known features (e.g. gene function and functional relationships) of the present profile data on a genomic scale, and a reliance on assumption of directed acyclic topology. Feedback loops are ubiquitous in biological processes and associated with many properties of gene networks, and thus analyses based on the assumption that no feedback loops exist are inappropriate.

Graphical Gaussian Model (GGM), also known as covariance selection model, was used as a model for the association network of genes. It can directly use the continuous expression profiling data without requiring other information. The partial correlation coefficients (PCCs) were used to characterize the strength of interaction between pair of genes and selection of partial correlations indicated by non-cyclic relationships among genes (Toh and Horimoto, 2002; Wang et al., 2003; Wu et al., 2003; Aburatani et al., 2005). However, the estimation procedure for statistical inference for individual treatments is not efficient.

Since the inferred network can be based on the partial correlation, calculating the matrix inverse of Pearson correlation coefficients is desirable. However, the number of genes to be analyzed usually far exceeds the number of expression measurements, and a high similarity in the expression pattern of some genes leads to strong collinearity among rows or columns in the correlation matrix. As a result, it is difficult to obtain the inverse of the correlation coefficient matrix, and is inappropriate to infer the regulatory relationships by simply using the partial correlation coefficient matrix (PCCM) to the expression profiles. To resolve this discrepancy, we propose a novel method based on a matrix of thresholding partial correlation coefficients (MTPCC). The idea is to use clusters, rather than individual genes, to eliminate the collinearity issue. There would be no linear relationship between any two clusters profiles. The regulatory network showed dependence among clusters as undirected dependency graph (UDG), its nodes and edges correspond to the clusters under consideration and direct interaction between clusters, respectively. The efficiency of our method was evaluated based on biological aspects by applying the method to the expression profiles of *Saccharomyces cerevisiae* (Eisen et al., 1998).

2. Materials and Methods

2.1. Gene Expression Profiles Data

Let $S=\{s_1, s_2, \ldots, s_m\}$ be the set of samples or conditions and $G=\{g_1, g_2, \ldots, g_n\}$ be the set of genes. The expression profiling data can be represented as $X=\{x_{ij}|i=1, \ldots, n, j=1, \ldots, m\}$ $(n \gg m)$, where x_{ij} corresponds to the expression value of the sample s_j on gene g_i. In order to evaluate our method, the gene expression data analyzed here are for $n=2467$ genes from *S. cerevisiae*, which were measured under $m=79$ conditions (Eisen et al., 1998) (http://www.pnas.org or http://rana.stanford.edu/clustering/). Missing data were estimated by using a k nearest neighbor method with an intermediate $(10 \le k \le 20)$ value of $k=14$ (Troyanskaya et al., 2001).

2.2. Procedure of MTPCC for Inferring the Regulatory Network

This procedure consists of three parts: (1) hierarchical cluster analysis, (2) cluster boundaries establishment, and (3) regulatory network inference.

2.2.1. Hierarchical Cluster Analysis

Hierarchical cluster analysis was performed to the gene expression data. The Pearson correlation coefficients of the standardized expression profiles were used for calculating distance, and the UPGMA method was applied for grouping genes. The $(n-1)$ dissimilarity scores of the nodes along the dendrogram were obtained by the hierarchical cluster analysis using ClusterProject (version ClusterProject1.0 from http://ibi.zju.edu.cn/software/clusterproject/, Pan et al., 2005).

2.2.2. Cluster Boundaries Establishment

Step 1: A correlation coefficient matrix (CCM) was obtained from the original CCM at each node along the dendrogram. For instance, when the dissimilarity score of the node $\hat{d}_c$ is set to be $\hat{d}_{2467-q+1} \ge \hat{d}_c > \hat{d}_{2467-q}$, q clusters are obtained and the $q \times q$ CCM is generated by the random selection of correlation coefficient from the gene members of each cluster.

Step 2: A statistical property of the $q \times q$ CCM obtained in Step 1 was evaluated along the dendrogram. The linear relationship between the clusters was diagnosed by the variance inflation factor (VIF), as follows:

$$\mathrm{VIF}_i = r_{ii}^{-1} \quad (1)$$

where r_{ii}^{-1} is the ith diagonal element of the inverse matrix of CCM. In a CCM for q clusters, q VIFs are calculated (Horimoto and Toh, 2001).

Step 3: In the diagnosis of extent of the linear relationship, the popular value of 10.0 was adopted as a threshold (Freund and Wilson, 1998). The q VIFs were evaluated under the following condition:

$$\max\{\mathrm{VIF}_i\} < 10.0 \quad \text{for } i = 1, 2, \ldots, q \quad (2)$$

If the condition (2) is satisfied, then there is no linear relationship among the q sets of clusters. Otherwise, the linear relationship still exists. The above steps from Step 1 to Step 3 proceed in a descending order of nodes from 2466 to 1, and the last node that satisfies the condition (2) is searched, so the maximum number of clusters with no linear relationship along the dendrogram is obtained.

2.2.3. Regulatory Network Inference

A network between the clusters obtained by the second part is inferred. The expression profiles of each cluster are expressed as average expression levels for the constituting genes of the cluster, and the number of conditions for the cluster are as same as the number of the measurement conditions, *i.e.* the expression level of the cluster k at the jth condition clu_k is calculated as follows:

$$\text{clu}_{kj} = \frac{1}{n_k} \sum_{i \in \text{cluster}\, k}^{n_k} x_{ij}, \quad 1 \leq k \leq n_1, 1 \leq j \leq m \tag{3}$$

where n_k is the number of members in the kth cluster and n_1 is the total number of obtained clusters.

When a set of expression levels of n_1 clusters is obtained under m different conditions, a PCCM can be calculated from the inverse of CCM for these clusters. For the PCCM $\mathbf{\Pi} = (\pi_{ij})$, these coefficients describe the correlation between any two clusters i and j conditioned on all the remainder of these clusters and are calculated as follows:

$$\pi_{ij} = -\frac{r_{ij}^{-1}}{\sqrt{r_{ii}^{-1} r_{jj}^{-1}}} \tag{4}$$

where r_{ij}^{-1}, r_{ii}^{-1} and r_{jj}^{-1} are the elements of the inverse of the $n_1 \times n_1$ correlation matrix $\mathbf{R}$.

If the value of π_{ij} is statistically indistinguishable from zero, then there is no detectable genetic link between clusters i and j. Finally, a graph of UDG, *i.e.* a regulatory network structure, which is visualized by Graphviz (Gansner and North, 2000), is obtained with its nodes and edges corresponding to the clusters and significant partial correlation coefficients, respectively.

In order to obtain the network, the significance of each element in the PCCM is inferred by the permutation test. We independently permute condition-profiles of each cluster, which are indexed from 1 to m. The profiles are shuffled by computing a random permutation of the indices $1, \ldots, m$ and assigning the ith expression data to the condition-profile whose index is given by the ith element of the permutation for each cluster. The shuffled sample data are then used to calculate a PCCM. This procedure is repeated k times, thus k PCCMs are obtained for the shuffled samples. Two types of threshold values can be estimated from these results. The first are comparison-wise thresholds that can be estimated separately for each element in the original PCCM, for example, the values of an element π_{ij} over the k PCCMs are sorted ascendingly, the estimated critical values are set as $100(1-\alpha/2)$ percentile and $100(\alpha/2)$ percentile. The test of using the critical values controls the type I error rate for that element to be α or less. The second are experiment-wise thresholds that can provide overall critical values for all analysis elements. They can be obtained by first finding the maximum and the minimum values over them in each PCCMs for the shuffled samples. Then the k maximum values are ordered, and their $100(1-\alpha/2)$ percentile is set as one of experiment-wise critical value. Similarly, another critical value is $100(\alpha/2)$ percentile of the k ordered minimum values. These critical values are used to control the overall type I error rate to be α or less. So the statistical significance of each element in the original PCCM can be obtained by comparing it with these critical values.

3. Results

3.1. Cluster Analysis

The 2467 yeast genes were classified into 12 clusters by hierarchical cluster analysis and cluster boundaries establishment (http://ibi.zju.edu.cn/lab/supplementary_materials_hanlide/). An unpaired t-test shows that the differences of the genes expression within each cluster (0.485 ± 0.199) are significantly ($p < 0.05$) smaller than those between the clusters (0.893 ± 0.212). Therefore we assumed that the genes in the same cluster share the same expression pattern, and that the expression levels of the cluster can represent the expression behavior of the constituted genes.

3.2. Regulatory Network Inference

The expression profiles of 12 clusters were permuted 1000 times. Comparison-wise thresholds were estimated for every element of these tests (not shown). The thresholds fluctuated across 66 elements and their average was listed in Table 1. The maximum and minimum partial correlation coefficients of all elements from each of the 1000 permutations were used to estimate the experiment-wise thresholds. The absolute values of comparison-wise thresholds are smaller than those of the corresponding experiment-wise thresholds and t critical values (Aburatani et al., 2003). This indicates that the comparison-wise thresholds are the least ones in terms of the evaluation of significance. In this example, we were interested in the comparison-wise thresholds. The significance level α was set as 0.05, and we obtained

Table 1
Estimated threshold value for expression data of *S. cerevisiae*

Threshold	$1-\alpha$	Experiment-wise	Comparsion-wise[a]	t critical value
+	0.95	0.377	0.195	0.237
−		−0.378	−0.196	−0.237
+	0.99	0.429	0.275	0.309
−		−0.429	−0.276	−0.309

"+" and "–" denote the critical value of the right and left tail, respectively.

[a] *Notes*: Average across all analysis elements.

Table 2
Partial correlation coefficient matrix obtained by permutation test

Cluster Name	1	2	3	4	5	6	7	8	9	10	11
1											
2	0.49**										
3	−0.03	0.23*									
4	−0.24*	0.29**	0.17								
5	0.31**	−0.30**	0.33**	−0.03							
6	−0.67**	0.55**	−0.11	−0.36**	0.13						
7	0.51**	−0.81**	0.02	0.47**	−0.02	0.74**					
8	0.32**	−0.36**	0.06	0.59**	0.13	0.67**	−0.67**				
9	0.09	0.07	−0.35**	0.07	0.06	0.22*	−0.14	−0.07			
10	0.18	0.45**	−0.14	0.18	0.11	−0.11	0.34**	−0.01	0.10		
11	0.07	0.45**	0.21*	−0.09	−0.14	−0.07	0.40**	0.07	0.14	−0.48**	
12	0.09	0.15	0.02	−0.09	0.19	0.10	0.10	0.04	−0.02	−0.08	0.08

Notes: The partial correlation coefficient of every pair of 12 clusters is shown. If the absolute value of the coefficients is bigger than the absolute comparison-wise threshold value, the corresponding element was considered as significant. * and ** denote statistical significance at the 0.05 and 0.01 levels, respectively. The insignificant elements in PCCM are shaded. The rows and the columns correspond to the clusters, and the cluster names are shown at the left and bottom of the matrix.

the symmetric PCCM (Table 2). The absolute values of the partial correlation coefficients ranged from 0.01 to 0.81. As the expression profiles of each cluster were expressed as a set of expression levels average over the constituted genes of the cluster, the signs of the coefficients did not always reflect the positive or negative regulations between these genes. Out of 66 coefficients, 39 (59.1%) were statistically insignificant from zero. In other words, 39 edges were removed from the graph of UDG. The graph did not contain any node without edges (Fig. 1). The maximum number of edges of a node was 9, while the minimum number was 2.

3.3. Regulatory Network Evaluation

With the inferred network through the method mentioned, we employed the previous published experimental literature to evaluate its validity. As a large amount of experimental data has been accumulated, it is impossible to collect all of the results of gene regulation in *S. cerevisiae*. We collected the related literatures and mainly focused on the regulation of *SUC2* (a gene for the sucrose hydrolyzing enzyme called invertase) expression as the gene has been investigated extensively. Under the assumption that the relationships defined by these experiments reflect the direct interactions about the genes expression, we evaluated the network with the results of the collected experimental studies. Forty-nine cases of regulatory relationships, which describe the relationship that Gene A affects the expression of Gene B (Table 3), were obtained from the related references in all. When the partial correlation coefficient between two clusters, corresponding to a pair of genes described in the literature, was significant, the inference of the relationship was regarded as being correct, otherwise the relationship was considered to be wrong. The estimations of these regulatory relationships by the MTPCC were shown in Table 3. In 5 out of 49 cases, both Genes A and B were present in the same cluster. The numbers of significant or insignificant partial correlation coefficients were 36 and 8 with the higher correct percentage (73.5%) and the lower false percentage (16.3%), respectively.

A regulatory network was obtained by a modified sub-graph of the UDG (Fig. 2) corresponding to the relationship depicted in experimental studies (Table 3). Each node corresponds to a cluster, and contains the genes that appear in Table 3, although only the genes related to *SUC2* expression are shown in Fig. 2. Both correct and incorrect relationships are included in the sub-graph.

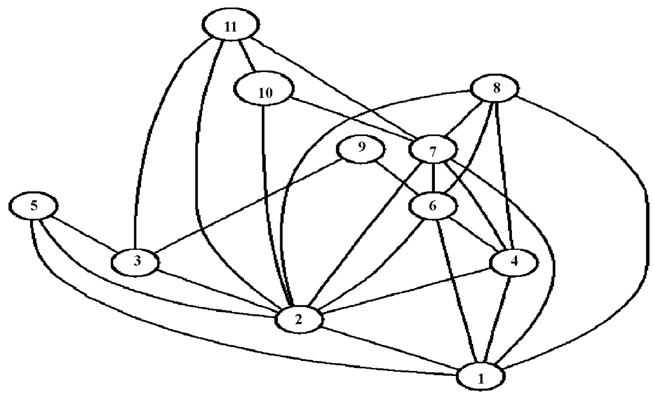

Fig. 1. A graph of UDG corresponding to the obtained PCCM (Table 2). A solid line indicates the interaction between a pair of clusters and the number in a node shows the name of the cluster.

Table 3
The relationships between genes for the regulation of expression

Gene A	CN	Gene B	CN	MTPCC	Bootstrap	Reference
SNF2	7	*ADH1*	2	T	F	Peterson and Herskowitz (1992) Cell 68, 573–583
SWI1	1			T	T	Peterson and Herskowitz (1992) Cell 68, 573–583
SNF2	7	*ADH2*	2	T	F	Peterson and Herskowitz (1992) Cell 68, 573–583
SWI1	1			T	T	Peterson and Herskowitz (1992) Cell 68, 573–583
SIN3	1	*BAR1*	4	T	F	Vidal et al. (1991) Mol. Cell Biol. 11, 6306–6316
SNF1	1	*MIG1*	2	T	T	Papamichos-Chronakis et al. (2004) EMBO Rep. 5, 368–372
RGR1	4	*CTS1*	9	F	F	Jiang et al. (1995) Genetics 140, 47–54
SIN4	10			F	F	Jiang et al. (1995) Genetics 140, 47–54
TUP1	8	*CYC1*	2	T	F	Zhang et al. (1991) Gene 97, 153–161
GAL11	1	*HIS4*	2	T	T	Sakurai et al. (1996) FEBS Lett. 398, 113–119
SIN4	10			T	T	Jiang and Stillman (1995) Genetics 140, 103–114
SNF2	7			T	F	Jiang and Stillman (1995) Genetics 140, 103–114
SFL1	1	*HSP26*	2	T	T	Lesage et al. (1994) Nucleic Acids Res. 22, 597–603
RGR1	4	*IME1*	10	F	T	Shimizu et al. (1998) Nucleic Acids Res. 26, 2329–2336
RME1	9			F	F	Shimizu et al. (1998) Nucleic Acids Res. 26, 2329–2336
SIN4	10			S	S	Shimizu et al. (1998) Nucleic Acids Res. 26, 2329–2336
HXK2	2	*MED8*	10	T	T	De la Cera et al. (2002) J. Mol. Biol. 319, 703–714
TUP1	8	*HIS2*	1	T	T	Watson et al. (2000) Genes Dev. 14, 2737–744
GAL11	1	*SUC2*	2	T	T	Vallier and Carlson (1991) Genetics 129, 675–684
GCN5	7			T	F	Pollard et al. (1999) EMBO J. 18, 5622–5633
RGR1	4			T	T	Sakai et al. (1988) Genetics 119, 499–506
ROX3	6			T	T	Song and Carlson (1998) EMBO J. 17, 5757–5765
SFL1	1			T	T	Song and Carlson (1998) EMBO J. 17, 5757–5765
SIN4	10			T	T	Song and Carlson (1998) EMBO J. 17, 5757–5765
SNF1	1			T	T	Neigeborn and Carlson (1984) Genetics 108, 845–858
SNF5	1			T	T	Neigeborn and Carlson (1984) Genetics 108, 845–858
SNF6	6			T	T	Neigeborn and Carlson (1984) Genetics 108, 845–858
SRB8	1			T	T	Song and Carlson (1998) EMBO J. 17, 5757–5765
SSN8	3			T	T	Kuchin et al. (1995) Proc. Natl. Acad. Sci. 92, 4006–4010
SWI1	1			T	T	Peterson and Herskowitz (1992) Cell 68, 573–583
SWI3	7			T	F	Peterson and Herskowitz (1992) Cell 68, 573–583
TUP1	8			T	F	Zhang et al. (2002) Genetics 161, 957–969
SSN3	2			S	S	Kuchin et al. (1995) Proc. Natl. Acad. Sci. 92, 4006–4010
SNF11	2			S	S	Peterson and Herskowitz (1992) Cell 68, 573–583
SNF2	7			T	F	Carlson and Laurent (1994) Curr. Opin. Cell Biol. 6, 396–402
MIG1	2			S	S	Trumbly (1992) Mol. Microbiol. 6, 15–21
HXK2	2	*MIG1*	2	S	S	Ahuatzi et al. (2007) J. Biol. Chem. 282, 4485–4493
MOT3	1	*ANB1*	8	T	T	Klinkenberg et al. (2005) Eukaryot. Cell 4, 649–660
MOT3	1	*HEM13*	6	T	F	Klinkenberg et al. (2005) Eukaryot. Cell 4, 649–660
TUP1	8	*RME1*	9	F	F	Mukai et al. (1991) Mol. Cell. Biol. 11, 3773–3779
SIN3	1	*RME1*	9	F	T	Vidal et al. (1991) Mol. Cell. Biol. 11, 6306–6316
TUP1	8	*SRB7*	6	T	T	Gromoller and Lehming (2000) EMBO J. 19, 6845–6852
HEM13	6	*ROX1*	1	T	F	Zhang et al. (2002) Genetics 161, 957–969
CYC8	10	*TUP1*	8	F	F	Zhang et al. (2002) Genetics 161, 957–969
TUP1	8	*RNR1*	7	T	F	Zhang et al. (2002) Genetics 161, 957–969
TUP1	8	*ROX1*	1	T	T	Mizuno et al. (1998) Curr. Genet. 33, 239–247
SIP3	3	*SNF1*	1	F	F	Conlan and Tzamarias (2001) J. Mol. Biol. 309, 1007–1015
CYC8	10	*SUC2*	2	T	T	Trumbly (1992) Mol. Microbiol. 6, 15–21
SNF4	2	*SNF1*	1	T	T	Shirra and Arndt (1999) Genetics 152, 73–87

Notes: The gene written in the first column (Gene A) is known to regulate the expression of the gene written in the third column of the same line (Gene B). The second and the fourth columns in the same line indicate the cluster names (CN), to which Genes A and B belong, respectively. The fifth and sixth columns include three symbols, 'T', 'F' and 'S'. A significant partial correlation coefficient between the corresponding clusters is regarded as accord with the experimental result, and 'T' is put in the column. An insignificant partial correlation coefficient between the corresponding clusters is regarded as being inconsistent with the experimental result, and 'F' is placed in the column. 'S' in the fifth and sixth columns indicate that both Genes A and B belong to the same cluster. The seventh column indicates the references for the experimental studies.

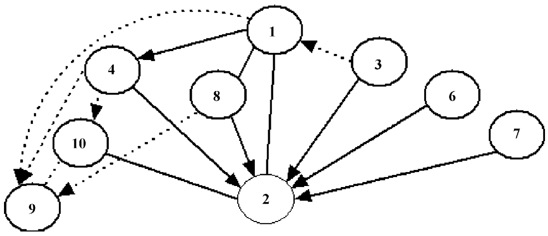

Fig. 2. A sub-graph of UDG corresponding to the relationship depicted in experimental studies (Table 3). A solid line indicates the interaction between a pair of clusters, which is also suggested by PCCM. Each node indicates a cluster. A dashed line indicates the regulatory relationship, which is not consistent with our inference. The arrows and the undirected edges indicate the cause and effect, feedback relationships suggested by the experimental results, respectively. The number in a node indicates the cluster name. The gene names within a cluster are written when they involved in the regulation of *SUC2* expression. 1: *SFL1*, *SNF1*, *SNF5*, *SRB8*, *SWI1*, *GAL11*; 2: *SUC2*, *SSN3*, *SNF11*, *MIG1*, *SNF4*; 3: *SSN8*; 4: *RGR1*, 6: *ROX3*, *SNF6*; 7: *GCN5*, *SNF2*, *SWI3*; 8: *TUP1*; 10: *SIN4*, *CYC8*.

SUC2 is included in cluster 2, its transcription activation depends upon both the *SNF1*/*SNF4* kinase complex and the *SWI*/*SNF* nucleosome complex (Zhou and Winston, 2001). *SNF1* and *SNF4* constitute the former, the later has 11 units, such as *SWI1*, *SWI2* (alias *SNF2*), *SWI3*, *SNF5*, *SNF6*, *SNF11*, etc. (Biggar and Crabtree, 1999; Carlson and Laurent, 1994). *SWI1*, *SNF4* and *SNF5* are included in cluster 1, cluster 2 contains *SNF11*, *SNF6* belongs to cluster 6 while cluster 7 includes *SNF2* and *SWI3*. As shown in Fig. 2, there are edges between cluster 2 and cluster 1, 6 and 7. *SUC2* expression is activated by *GCN5* (Pollard et al., 1999), and regulated negatively by *SSN3* and *SSN8* (Kuchin et al., 1995). *GCN5*, *SSN3* and *SSN8* belong to clusters 7, 2 and 3, respectively. The edges between clusters 2 and 3, 7 are present. Mutations of *SRB8*, *SIN4* or *ROX3* cause *SUC2* the defect in transcriptional repression, which can be suppressed by *SFL1* gene (Song and Carlson, 1998). *SRB8* and *SFL1* belong to cluster 1, *SIN4* is included in cluster 10, and *ROX3* is involved in cluster 6. The presence of edges between cluster 2 and clusters 1, 10 and 6 supports this observation. *SIN4*, *RGR1*, *GAL11* and *p50* form a regulatory sub-complex to control transcription (Li et al., 1995). In addition, *TUP1*, *CYC8* and *MIG1* are regarded as a complex for regulation of glucose repression related genes (Tzamarias and Struhl, 1994), *GAL11* is included in cluster 1, cluster 10 contains *SIN4* and *CYC8*, *RGR1* belongs to cluster 4, *TUP1* is involved in cluster 8, *MIG1* and *SUC2* are in same cluster. These interactions were indicated by the edges between clusters 1, 4, 8, 10 and cluster 2. Thus, the collected experimental studies results regarding *SUC2* regulation are consistent with the edges in the UDG. Similarity, most of the remaining edges also accord with the other collected expression regulatory relationships. The coincidence of our results with reported experimental studies indicated that our method was effective.

The obtained graph of UDG is basically undirected. According to the causality relationships obtained from the literatures, the edges were replaced with arrows indicating the causes and effects. Each arrow in the graph indicated plural of regulatory relationships (Fig. 2). For example, the arrow which connecting cluster 7 with cluster 2 corresponds to the relationships between six gene pairs. A loop relationship was also observed between two clusters, such as the edges connected cluster 1 with cluster 2 correspond to the feedback relationships, which implied that a subset of genes within cluster 2 directly affected the expression of ones within cluster 1 and vice versa. For example, *SNF4* (cluster 2) regulates *SNF1* (cluster 1), while *SWI1* (cluster 1) affects *SUC2* (cluster 2).

4. Discussions

The statistical parametric tests for network inference, such as a *t*-test, are very powerful tools when the data follow a particular distribution. But they are less suitable than some other methods, for instance GGM, when applied to expression data (Aburatani et al., 2003). In contrast, nonparametric tests make less stringent demands of the data. Therefore, a permutation test is an appropriate method for distinguishing the significant regulatory relationship of genes. The sufficient shuffling replications of sample are also important for the test, and shuffling 1000 times is considered to be appropriate to give some critical values with $\alpha = 0.05$. The permutation tests are much more powerful than bootstrapping when they are used to construct a test of a hypothesis of edges. We found 28 edges (http://ibi.zju.edu.cn/lab/supplementary_materials_hanlide/) when bootstrapping was applied to the expression profiles of *S. cerevisiae*, while only 18 of these matched those found by a permutation test. Moreover, the bootstrap method suffered a lower correct percentage (53.1%) and a higher false percentage (36.7%) from the biological viewpoint (Table 3). In addition, the method of combining bootstrap sample with GGM was not sufficiently effective because relatively high bootstrap probabilities were sometimes observed even at the insignificant elements in the original PCCM (Toh and Horimoto, 2002).

In this paper we presented an effective approach for inferring regulatory network from gene expression profiles, and the approach can be regarded as an extension of the works of some researcher (Toh and Horimoto,

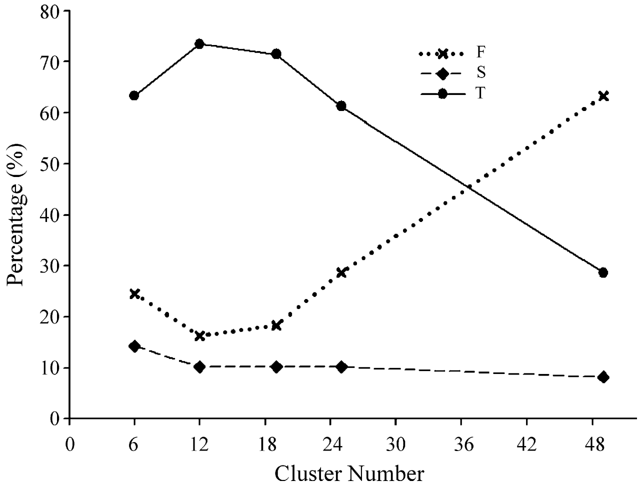

Fig. 3. Situation that the literature fits network inferred by MTPCC under different cluster number. 'T', 'F' and 'S' were defined as in Table 3.

2002; Wang et al., 2003; Wu et al., 2003; Aburatani et al., 2005). GGM is a suitable method for network inference, but it assumes the observed data following a multivariate normal distribution. In fact, the gene expression levels are often non-normally distributed and do not match the assumption. The estimation of the method is most likely obtained by chance but not reflecting the truth. MTPCC method does not require the data following any specific statistical distribution. It is valid under very mild conditions and easy to apply in practice. When applying GGM to expression data of *S. cerevisiae*, 29 edges were found (http://ibi.zju.edu.cn/lab/supplementary_materials_hanlide/), while 25 were also detected by MTPCC. From biological aspect, the network inferred by MTPCC explained some regulation relationships about *SUC2* and other genes of *S. cerevisiae* with high correct percentage and low false percentage. As seen in Fig. 3, cluster number influences the power of our method. Under the condition of 12 clusters (VIF = 10.0), MTPCC infers the network with higher correct percentage and lower false percentage. The correctness demonstrated its accuracy and efficiency, thus MTPCC is a valid statistical approach for inferring the regulatory network.

Acknowledgements

We greatly thank the anonymous reviewer for useful comments and suggestions on the earlier version of the manuscript. This work was partially supported by the National Basic Research Program of China (973 Program) and the National Natural Science Foundation of China. We are grateful to Yangyun Zou, Xusheng Wang, Guobo Chen, and Yousaf Hayad for their careful reading of this manuscript.

References

Aburatani, S., Goto, K., Saito, S., Toh, H., Horimoto, K., 2005. ASIAN: a web server for inferring a regulatory network framework from gene expression profiles. Nucleic Acids Res. 33, 659–664.

Aburatani, S., Kuhara, S., Toh, H., Horimoto, K., 2003. Deduction of a gene regulatory relationship framework from gene expression data by the application of graphical Gaussian modeling. Signal Process. 83, 777–788.

Akutsu, T., Miyano, S., Kuhara, S., 1999. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. In: Proceedings of the Pacific Symposium on Biocomputing.

Biggar, S.R., Crabtree, G.R., 1999. Continuous and widespread roles for the *Swi–Snf* complex in transcription. EMBO J. 18, 2254–2264.

Carlson, M., Laurent, B.C., 1994. The *SNF/SWI* family of global transcriptional activators. Curr. Opin. Cell Biol. 6, 396–402.

DeRisi, J.L., Vishwanath, R.I., Patrick, O., 1997. Exploring the metabolic genetic control of gene expression on a genomic scale. Science 278, 680–686.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95, 14863–14868.

Freund, R.J., Wilson, W.J., 1998. Regression Analysis. Academic Press, San Diego.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7, 601–620.

Gansner, E.R., North, S.C., 2000. An open graph visualization system and its applications to software engineering. Softw. Pract. Exp. 30, 1203–1233.

Hartemink, A.J., Gifford, D.K., Jaakkola, T., Young, R.A., 2002. Combining location and expression data for principled discovery of genetic regulatory network models. Pac. Symp. Biocomput. 7, 437–449.

Horimoto, K., Toh, H., 2001. Statistical estimation of cluster boundaries in gene expression profile data. Bioinformatics 17, 1143–1151.

Kuchin, S., Yeghiayan, P., Carlson, M., 1995. Cyclin-dependent protein kinase and cyclin homologs *SSN3* and *SSN8* contribute to transcriptional control in yeast. Proc. Natl. Acad. Sci. 92, 4006–4010.

Li, Y., Bjorklund, S., Jiang, Y.W., Kim, Y.J., Lane, W.S., Stillman, D.J., Kornberg, R.D., 1995. Yeast global transcriptional regulators *Sin4* and *Rgr1* are components of mediator complex/RNA polymerase II holoenzyme. Proc. Natl. Acad. Sci. 92, 10864–10868.

Murphy, K., Mian, S., 1999. Modeling gene expression data using dynamic Bayesian networks. Technical report. Computer Science Division, University of California, Berkeley, CA.

Pan, H.Y., Zhu, J., Han, D.F., 2005. Clustering gene expression data based on predicted differential effects of GV interaction. Genomics Proteomics Bioinformatics 3, 36–41.

Pollard, K., Samuels, M.L., Crowley, K.A., Hansen, J.C., Peterson, C.P., 1999. Functional interaction between *GCN5* and polyamines: a new role for core histone acetylation. EMBO J. 18, 5622–5633.

Somogyi, R., Shiegoski, C.A., 1996. Modeling the complexity of genetic networks: understanding multigene and pleiotropic regulation. Complexity 1, 45–63.

Song, W., Carlson, M., 1998. Srb/mediator proteins interact functionally and physically with transcriptional repressor Sfl1. EMBO J. 17, 5757–5765.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Toh, H., Horimoto, K., 2002. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. Bioinformatics 18, 287–297.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics 17, 520–525.

Tzamarias, D., Struhl, K., 1994. Functional dissection of the yeast *Cyc8–Tup1* transcriptional co-repressor complex. Nature 369, 758–761.

Wang, J., Myklebost, O., Hovig, E., 2003. MGraph: graphical models for microarray data analysis. Bioinformatics 19, 2210–2211.

Wu, X.T., Ye, Y., Subramanian, K.R., 2003. Interactive analysis of gene interactions using graphical Gaussian model. BIOKDD 3, 63–69.

Zhou, H., Winston, F., 2001. NRG1 is required for glucose repression of the *SUC2* and *GAL* genes of *Saccharomyces cerevisiae*. BMC Genet. 2, 5.