



Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling

Hiroyuki Toh^{1,*} and Katsuhisa Horimoto^{2,†}

¹Department of Bioinformatics, Biomolecular Engineering Research Institute 6-2-3, Furuedai, Suita, Osaka 565-0874, Japan and ²Laboratory of Mathematics, Saga Medical School, 5-1-1 Nabeshima, Saga, Saga 849-8501, Japan

Received on June 29, 2001; revised and accepted on August 9, 2001

ABSTRACT

Motivation: Recent advances in DNA microarray technologies have made it possible to measure the expression levels of thousands of genes simultaneously under different conditions. The data obtained by microarray analyses are called expression profile data. One type of important information underlying the expression profile data is the 'genetic network,' that is, the regulatory network among genes. Graphical Gaussian Modeling (GGM) is a widely utilized method to infer or test relationships among a plural of variables.

Results: In this study, we developed a method combining the cluster analysis with GGM for the inference of the genetic network from the expression profile data. The expression profile data of 2467 *Saccharomyces cerevisiae* genes measured under 79 different conditions (Eisen *et al.*, *Proc. Natl Acad. Sci. USA*, **95**, 14 683–14 868, 1998) were used for this study. At first, the 2467 genes were classified into 34 clusters by a cluster analysis, as a preprocessing for GGM. Then, the expression levels of the genes in each cluster were averaged for each condition. The averaged expression profile data of 34 clusters were subjected to GGM, and a partial correlation coefficient matrix was obtained as a model of the genetic network of *S. cerevisiae*. The accuracy of the inferred network was examined by the agreement of our results with the cumulative results of experimental studies.

Availability: A set of programs will be electronically sent upon request.

Contact: toh@beri.co.jp; horimoto@ged.saga-med.ac.jp

INTRODUCTION

Recent advances in DNA microarray technologies have made it possible to measure the expression levels of thousands of genes simultaneously, under different conditions.

Here, different conditions mean different stages of the cell cycle, different developmental stages, different body tissues, different clinical conditions or different organisms. Some conditions can be ordered. For example, an order in terms of time progress can be assigned to the conditions related to the cell cycle or developmental stage. A set of expression levels of genes measured under various conditions is referred to here as the expression profile of the genes.

Elucidating patterns from the expression profile would provide great insight into gene function and regulatory systems. For this purpose, several groups have developed methods for clustering genes on a microarray. Here, clustering means partitioning the genes on the microarray into distinctive sets of genes that show similar expression patterns across the conditions. Hierarchical clustering (Eisen *et al.*, 1998), self-organizing mapping (Tamayo *et al.*, 1999), and other clustering methods (Ben-Dor *et al.*, 1999) have been applied to the expression profile data. Clustering genes with expression profiles can be utilized to predict the functions of gene products with functions that are unknown, and to identify sets of genes that are regulated by the same mechanism.

One of the important types of information underlying the expression profile data is the regulatory networks among genes, which here is called the 'genetic network.' Modelings with the Boolean network (Somogyi and Shiegoski, 1996), differential equations (Chen *et al.*, 1999; D'haeseleer *et al.*, 1999), and a combination of the methods (Akutsu *et al.*, 2000) have been investigated for inferences of the genetic networks. Tavazoie *et al.* (1999) proposed an approach that combines cluster analysis with sequence motif detection, to determine the genetic network architecture. Recently, an approach to infer the genetic networks with Bayesian networks was proposed (Friedman *et al.*, 2000).

In this paper, we describe a novel approach to infer the genetic networks from the expression profiles. In our ap-

*To whom correspondence should be addressed.

† Both authors contributed equally to this work.

proach, the genetic networks are inferred by a combination of cluster analysis and a method called 'Graphical Gaussian Modeling' (GGM; Whittaker, 1990; Edwards, 1995). We performed the cluster analysis as a preprocessing of the expression data at first. Then, the expression profiles for the obtained clusters were subjected to GGM. The efficiency of our approach was examined by the application of our method to the expression profile data of *Saccharomyces cerevisiae* (Eisen *et al.*, 1998). The accuracy of the inferred network will be discussed from both biological and statistical viewpoints.

MATERIALS AND METHODS

Expression profile data

Suppose that the number of genes under examination is L . Let N be the number of different conditions under which the expression levels of the L genes are measured. Then, the expression profile of the genes is represented as a set of L -dimensional vectors ($\text{el}(\text{Gene } 1(j)), \text{el}(\text{Gene } 2(j)), \dots, \text{el}(\text{Gene } L(j))$), where $\text{el}(\text{Gene } i(j))$ indicates the expression level of Gene i under the condition j ($1 \leq i \leq L, 1 \leq j \leq N$). To examine our method, we used the expression profile data for 2467 genes from *S. cerevisiae*, which were measured under 79 different conditions (Eisen *et al.*, 1998). That is, $L = 2467$ and $N = 79$ in this case. The data were obtained from a web site, <http://rana.stanford.edu/clustering/>. Let $\mu(i)$ be the average of the expression level of the i th gene over N different conditions. Then, the covariance, $s(i, j)$, between the i th and j th genes is defined as $(1/N) \sum_{k=1}^N (\text{el}(\text{Gene } i(k)) - \mu(i))(\text{el}(\text{Gene } j(k)) - \mu(j))$. The covariance matrix of the genes, Σ , is an $L \times L$ symmetric matrix. Let $\sigma(i)$ be the standard deviation of the expression level of the i th gene over N different conditions. A correlation coefficient matrix is a symmetric matrix, C , derived from Σ , whose elements are expressed as $c(i, j) = s(i, j)/(\sigma(i)\sigma(j))$, $1 \leq i, j \leq L$.

Cluster analysis as a preprocessing for GGM

As described below, the calculation of the inverse of the covariance matrix Σ is required for GGM. However, the results of the cluster analyses of the expression profile data by Eisen *et al.* (1998) suggest that many genes share similar expression patterns. A high similarity in the expression pattern induces a linear dependence among rows or columns in the correlation coefficient matrix, in terms of numerical analysis, which makes the calculation of the inverse matrix difficult. Therefore, we performed a hierarchical cluster analysis to classify the 2467 genes into some clusters, so that no linear dependence would be observed among the representative genes of the clusters in the correlation coefficient matrix. The details of the cluster analysis are discussed in the accompanying paper

(Horimoto and Toh, 2001). As a result of the cluster analysis, the 2467 genes were reduced to 34 clusters (see below). After the cluster analysis, the expression levels of the genes of a cluster were averaged for each condition of the measurement. Then, the expression profile of a cluster was expressed as a set of the expression levels averaged over the constituent genes of the cluster, and the size of the set was the same as the number of conditions for measurement. The procedure is defined as follows. Let M be the total number of obtained clusters. In the current application, $M = 34$. Suppose that a cluster k includes n genes. Consider the j th condition for measurement. Then, the averaged expression level of the cluster k at the j th condition is calculated as follows:

$$\text{el}(\text{cluster } k(j)) = (\sum_{\text{Gene } i \in \text{cluster } k} \text{el}(\text{Gene } i(j))) / n, \\ 1 \leq k \leq M, 1 \leq j \leq N.$$

Instead of the expression profile of genes, the expression profile of the clusters is hereafter considered, which is represented as a set of M -dimensional vectors ($\text{el}(\text{cluster } 1(j)), \text{el}(\text{cluster } 2(j)), \dots, \text{el}(\text{cluster } M(j))$), where $\text{el}(\text{cluster } i(j))$ indicates the averaged expression level of cluster i under condition j ($1 \leq i \leq M, 1 \leq j \leq N$). The averaged profile data were subjected to the analysis by GGM.

CONCEPTUAL FRAMEWORK OF GGM FOR INFERENCE OF GENETIC NETWORKS

The correlation coefficient in the expression profile data has been widely utilized to evaluate the distance between genes for the cluster analysis (Eisen *et al.*, 1998). Suppose that a pair of genes, say Genes A and B, show a high correlation in their expression profiles. There are three possible mechanisms to induce a high correlation in the expression levels between them. The first is a direct interaction between the genes. The second is an indirect interaction between them. In other words, the regulatory information of the Gene A product is transferred through the expressions of some other genes to induce the expression of Gene B. The third is the correlation due to the regulation by a common gene. That is, the expressions of Genes A and B are regulated by a common gene product. A combination of the second and third types of interactions would also cause a high correlation between the genes. The first type of interaction is what we want to know in order to reconstruct the genetic network from the expression profile data, although a correlation coefficient cannot distinguish between the three types of interactions. GGM is a multivariate analysis to infer or test a statistical model for the relationship among a plural of variables (Whittaker, 1990; Edwards, 1995), where a partial correlation coefficient, instead of a correlation coefficient, is used as a measure to select the

first type of interaction. In GGM, the statistical model for the relationship among the variables is represented as a graph, called the ‘independence graph,’ where the nodes correspond to the variables under consideration, and the edges correspond to the first type of interactions between variables. More specifically, an edge in the independence graph indicates a pair of variables that are conditionally dependent.

In GGM, the relationship among variables is inferred from a set of observed data of the variables, guided by the idea of conditional independence (Whittaker, 1990; Edwards, 1995). Due to the reason described above, a simple application of GGM to the expression profiles failed to infer the genetic networks. Therefore, the averaged expression levels of clusters are hereafter considered, instead of the raw gene expression profile data. Suppose that we have a data set of averaged expression levels of M clusters measured under N different conditions, each of which is represented as an M -dimensional vector ($\text{el}(\text{cluster } 1(i)), \text{el}(\text{cluster } 2(i)), \dots, \text{el}(\text{cluster } M(i))$) and $1 \leq i \leq N$. The averaged profile data set was constructed by the procedure described above. In order to apply GGM, we assumed that each of the M -dimensional vectors is drawn from a multivariate normal distribution. Then, the conditional distribution of the expression levels of any pair of clusters i and j , given the expression levels of $M - 2$ clusters, is a 2-dimensional normal distribution. One of the parameters of the conditional distribution is a covariance matrix of the averaged expression levels of clusters i and j under the condition that the averaged expression levels of the remaining genes are given. The elements of the covariance matrix are calculated with the elements of the inverse of the original $M \times M$ covariance matrix. Let $\Omega(\omega^{ij})$ be the inverse covariance matrix or the precision matrix, Σ^{-1} . Then, the diagonal elements of the 2-dimensional conditional covariance matrix are given as ω^{ii} and ω^{jj} , and the off-diagonal element is given as ω^{ij} . If $\omega^{ij} = 0$, then the conditional normal distribution is expressed as the product of the function of the averaged expression level of cluster i and that of cluster j . That is, clusters i and j are conditionally independent in their expression levels, given the remaining $M - 2$ clusters’ averaged expression levels, when $\omega^{ij} = 0$. In the application of GGM, conditional independence between a pair of variables i and j is evaluated using the partial correlation coefficient between the variables, $\rho^{ij, \text{the rest}}$, instead of ω^{ij} (Whittaker, 1990; Edwards, 1995), and $\rho^{ij, \text{the rest}}$ is given as $-\omega^{ij}/(\omega^{ii} \times \omega^{jj})$. That is, the variables i and j are conditionally independent when $\rho^{ij, \text{the rest}} = 0$.

GGM procedure

We have explained the conceptual framework of GGM, using the relationship between a pair of clusters as an example. In the actual application of GGM, however, we

should identify a set of cluster pairs that are conditionally independent. In this section, we will explain the procedure known as the Wermuth and Scheidt algorithm (1977). A graph, $G = (V, E)$, was used to represent the relationship among the M clusters, where V is a finite set of nodes, each corresponding to the M clusters, and E is a finite set of edges between the nodes. E consists of the edges between cluster pairs with averaged expression levels that are conditionally dependent, given the rest. The graph G expresses the Markov property among the variables. In other words, any pair of variables, which are not connected in the graph, is conditionally independent. In order to evaluate which pair of clusters is conditionally independent, we applied a stepwise and iterative algorithm developed by Wermuth and Scheidt (1977).

- Step 0.* A complete graph of $G(0) = (V, E)$ is prepared. The nodes correspond to M clusters. All of the nodes are connected. $G(0)$ is called a full model. Based on the expression profile data, an initial correlation coefficient matrix $C(0)$ is constructed.
- Step 1.* Calculate the partial correlation coefficient matrix (PCCM) $P(\tau)$ from the correlation coefficient matrix $C(\tau)$. τ indicates the number of the iteration.
- Step 2.* Find an element that has the smallest absolute value among all of the non-zero elements of $P(\tau)$. Then, replace the element in $P(\tau)$ with zero.
- Step 3.* Reconstruct the correlation coefficient matrix, $C(\tau + 1)$, from $P(\tau)$. In $C(\tau + 1)$, the element corresponding to the element set to zero in $P(\tau)$ is revised, while all of the other elements are left to be the same as those of $C(\tau)$.
- Step 4.* In the Wermuth and Scheidt algorithm, the termination of the iteration is judged by the values called ‘deviance.’ Here we used two types of deviance, dev 1 and dev 2, with the following definitions:

$$\text{dev 1} = N \log(|C(\tau + 1)|/|C(0)|),$$

$$\text{dev 2} = N \log(|C(\tau + 1)|/|C(\tau)|).$$

Calculate dev 1 and dev 2.

The two deviances follow an asymptotic χ^2 -distribution with a degree of freedom = n , and that with a degree of freedom = 1, respectively. n is the number of elements that are set to zero until the $(\tau + 1)$ th iteration. In our approach, n is equal to $(\tau + 1)$. $|C(\tau)|$ indicates the determinant of $C(\tau)$. N is the number of different conditions under which the expression levels of M clusters are measured.

- Step 5.* If the probability value corresponding to $\text{dev 1} \leq 0.05$, or the probability value corresponding to $\text{dev 2} \leq 0.05$, then the model $C(\tau + 1)$

is rejected, and the iteration is stopped. Otherwise, the edge between a pair of clusters whose partial correlation coefficient is set to zero in $P(\tau)$ is omitted from $G(\tau)$ to generate $G(\tau + 1)$, and τ is increased by 1. Then, go to *Step 1*.

The graph obtained by this procedure is an undirected graph, which is called an independence graph. The independence graph represents which pair of clusters is conditionally independent. That is, when the partial correlation coefficient for a cluster pair is equal to 0, the cluster pair is conditionally independent, and the relationship is expressed as no edge between the nodes corresponding to the clusters in the independence graph. In essence, the graph represents the genetic network of the M clusters under consideration.

RESULTS

Cluster analysis

The 2467 *S. cerevisiae* genes (Eisen *et al.*, 1998) were classified into 34 clusters by a hierarchical cluster analysis (Horimoto and Toh, 2001). There is not enough space here to describe the gene compositions of the 34 clusters. The details of the cluster analysis can be found in a web site (see <http://www.beri.co.jp/~protein/> or <http://www.saga-med.ac.jp/horimoto/microarray/>).

Further expression profile data of *S. cerevisiae* genes would make it possible to classify the 2467 genes into more clusters including smaller numbers of genes. The genes belonging to the same cluster could not be distinguished in terms of their expression levels from each other. Therefore, we assumed that the expressions of the genes in the same cluster are regulated in a similar manner, and that the expression levels averaged over the genes of each cluster can represent the expression behavior of the cluster.

Graphical Gaussian modeling

The iterative procedure of GGM was stopped when the probability value for either dev 1 or dev 2 did not satisfy the given level of significance. Here, the probability values for dev 1 and dev 2 at each iteration step were plotted as a function of the step number (see Figure 1). As shown in the figure, the probability values for dev 1 were always greater than 99.999% over the iterative calculation. In contrast, those of dev 2 gradually decreased with increasing step numbers. Finally, the iterative calculation was stopped at the step number = 189. The probability value for dev 2 at the 189th step was 0.011. Therefore, the number of iteration steps before stopping the procedure, 188, corresponded to the number of elements of PCCM that were replaced with 0.0. As shown in Figure 1, dev 2 was effective for the judgment of stopping the iteration, while dev 1 was not a good measure for the judgment.

The lower half of PCCM for the 34 clusters is shown

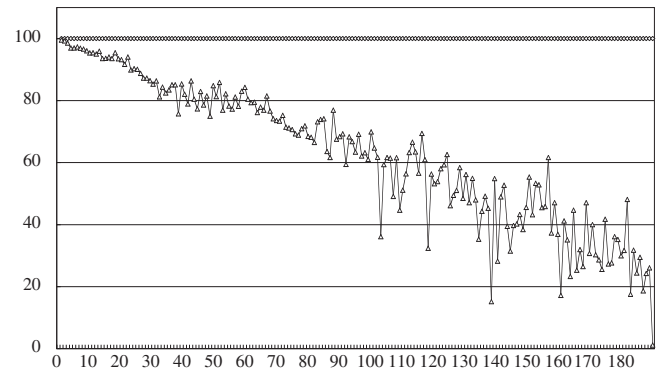


Fig. 1. Plots of probability values for dev 1 and dev 2 as a function of iteration step number. The probability values for dev 1 and dev 2, used for the judgment of stopping the iterative procedure for GGM, are plotted as a function of the step number of the iteration. The scale of the vertical axis is the probability value $\times 100$, while the horizontal axis indicates the iteration step number. Lozenges and triangles indicate the plots of the probability values for dev 1 and dev 2, respectively.

in Figure 2, since the matrix is symmetric. An element (33, 28) had the lowest partial correlation coefficient ($=-0.61$) within the matrix, while another element (33, 32) had the largest coefficient ($=0.78$). Out of 561 elements, 188 (about 34%) were replaced with 0.0 by the iterative procedure of GGM. In other words, 188 edges were removed from the independence graph. The independence graph did not include any node without edges. Inversely, there was no node with edges to all of the other nodes in the graph. The maximum number of edges of a node was 31, while the minimum number was 17. Since many edges remained, it would be quite difficult to show all of the edges in the independence graph. A modified subgraph of the obtained independence graph is shown in Figure 3. Details of the subgraph are discussed in the next section.

Hereafter, we will not mention the values of the partial correlation coefficients, because the signs of the values did not always reflect the positive or negative regulations experimentally observed. This would be caused by the averaging of the expression levels in the same clusters. Instead, we discuss the result of our method, and focus on the zero or non-zero features of the elements of the obtained PCCMs. In other words, we will examine the accuracy of our approach, based only on presence or absence of edges in the independence graph. In this manuscript, the presence of an edge in the independence graph is used with the same meaning as a non-zero partial correlation coefficient in PCCM.

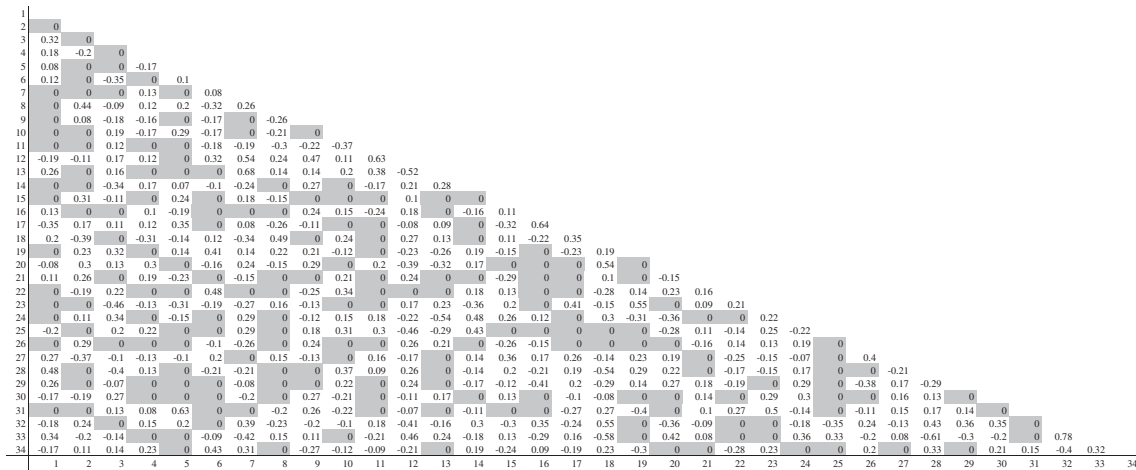


Fig. 2. Partial correlation coefficient matrix obtained by GGM. The partial correlation coefficient of every pair of 34 clusters is shown, where the elements replaced with 0.0 in the iterative procedure of GGM are shaded. The rows or columns correspond to the clusters, and the cluster numbers are shown at the left and bottom of the matrix.

DISCUSSION

Evaluation of the inferred genetic network from biological viewpoints

Does an edge in the inferred independence graph, or a non-zero element in the obtained PCCM, indicate a regulatory relationship about the expression between the genes included in the clusters connected by the edge? To check this problem, we examined the correspondence of the edges with regulatory relationships directly or indirectly suggested by experimental studies. However, it was difficult to collect all of the literature of the experimental studies of the regulation of expression in *S. cerevisiae*, because a large amount of experimental data have been accumulated. Therefore, we collected the literature, that focused on the regulation of SUC2 expression, because the SUC2 gene has been extensively studied. Some literature about the expression of different genes from SUC2 was found during the collection process, and we also used it to examine our result. The references are listed in Table 1. Then, we evaluated the obtained PCCM with the results of the collected experimental studies, under the assumption that the relationships defined by the experiments reflect the direct interactions about the expression of the genes. We obtained 40 cases of the regulatory relationships, which describe the relationship that Gene A affects the expression of Gene B (see Table 1). When the partial correlation coefficient between two clusters, corresponding to a pair of genes described in the literature, was not zero, the inference of the relationship was regarded as being correct. Otherwise, the relationship inferred by GGM was considered to be wrong. The results are summarized in Table 1. In 3 out of 40 cases,

both Genes A and B were present in the same cluster. On the other hand, the corresponding partial correlation coefficients were 0.0 in 8 out of the 40 cases. However, the remaining 29 pairs of clusters had non-zero partial correlation coefficients.

A subgraph of the independence graph corresponding to Table 1 is shown in Figure 3. Each node corresponds to a cluster, and includes the genes that appear in Table 1, although only the genes related to SUC2 expression are written in the nodes. Both correct and incorrect relationships are included in the subgraph. SUC2 is a gene for the sucrose hydrolyzing enzyme called invertase, which is included in cluster 23. SNF1–3 are considered to constitute a large complex together with SWI1 and SWI3, to form a supermolecule involved in the expression regulation of various genes, including SUC2 (Peterson and Herskowitz, 1992). Cluster 9 included SWI1, SNF2, and SWI3 (SNF2 is another name of SWI2). SNF5 is included in cluster 26, while cluster 5 contains SNF6. As shown in Figure 3 and Table 1, there are edges between cluster 23 and clusters 9, 5, and 26. In other words, the partial correlation coefficients corresponding to the edges were not zero. GAL11 (also known as SPT13 and RAP3) has been identified as a transcriptional regulator of galactose metabolizing enzymes, but the gene is also involved in the regulation of SUC2. GAL11 is included in cluster 26, as well as SNF5. That is, the interaction was indicated by the edge between clusters 23 and 26. SIN4 (also known as BEL2) and RGR1 are considered to form a complex for transcriptional regulation (Li *et al.*, 1995). SIN4 is included in cluster 24, while RGR1 belongs to cluster 28. Both of them are involved in the

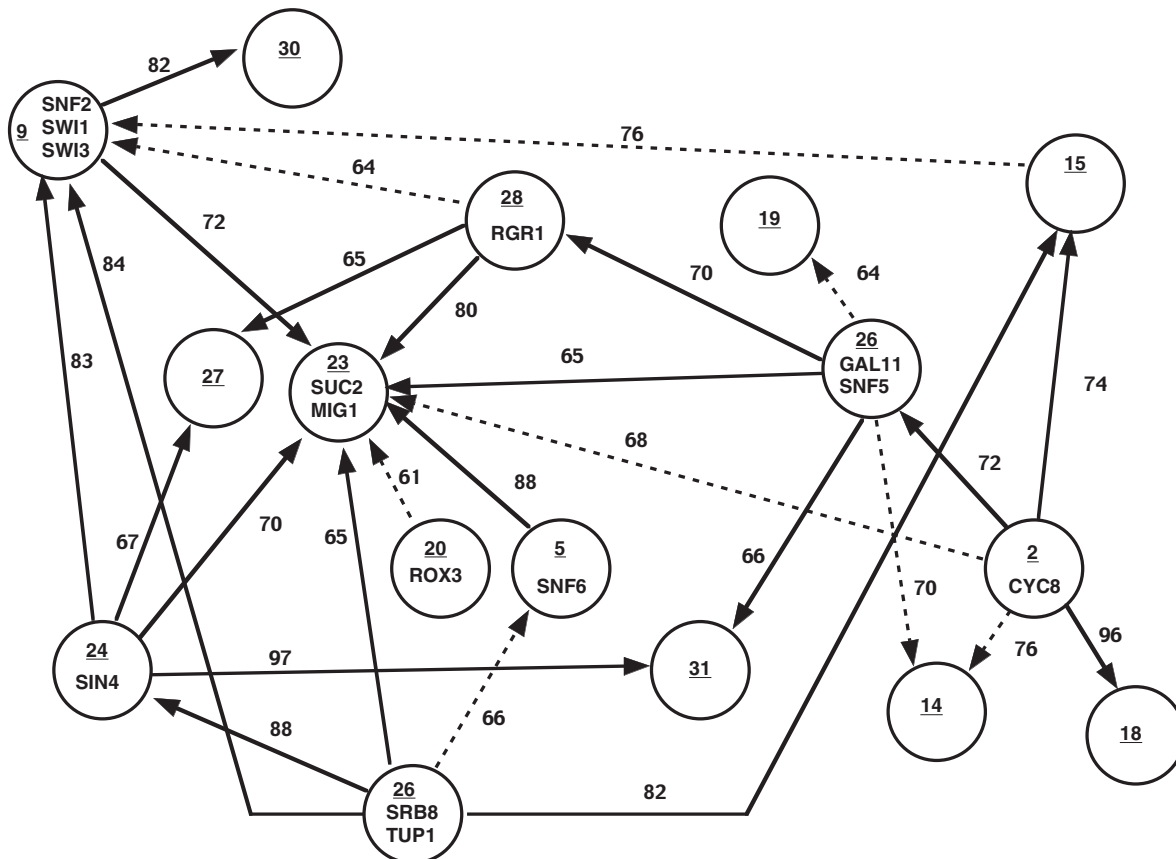


Fig. 3. A subgraph of the independence graph corresponding to the obtained partial correlation coefficient matrix. The clusters listed in Table 1 and the relationships among the clusters are shown in the figure. A solid line indicates the interaction between a pair of clusters, which are also suggested by our approach. Each node indicates a cluster. A dashed line indicates the regulatory relationship, which disagrees with our inference. The edges of the independence graph are basically obtained as undirected edges. However, the edges were replaced with arrows, according to the causes and results suggested by the experimental results (see Table 1). The underlined number in a node indicates the identification number of the cluster, and the number associated with each edge indicates the bootstrap probability for the edge between a pair of clusters. The gene names of the members of a cluster are written when they are involved in the regulation of SUC2 expression.

regulation of SUC2 expression. The presence of edges between cluster 23 and clusters 24 and 28 supports this observation. TUP1, CYC8, and MIG1 are also considered to form a complex for the regulation of glucose repression-related genes (Tzamarias and Struhl, 1994). TUP1 belongs to cluster 26. CYC8 (also known as SSN6) is included in cluster 2. MIG1 is included in cluster 23, along with SUC2. The edge between clusters 23 and 26 is present, but there is no edge between clusters 2 and 23, since the corresponding partial correlation coefficient was zero (see Figure 1). SRB8 is involved in the SUC2 expression, and belongs to cluster 26, like TUP1. Thus, most of the collected experimental studies about the regulation of SUC2 are consistent with the GGM results. Likewise, most of the remaining edges are consistent with the other collected expression regulatory relationships besides those

of SUC2 (see Table 1). Thus, the results suggest that our approach can explain the experimental studies of the expression regulation to some extent, although the genetic network is inferred as the relationships among the clusters of genes.

As described above, the graph obtained by GGM is basically undirected. According to the causality relationships obtained from the literatures, however, the edges were replaced with arrows to indicate the causes and the results. As shown in Figure 3, the obtained subgraph is acyclic. Each arrow in the graph indicates plural of regulatory relationships. For example, the edge connecting cluster 9 with cluster 30 corresponds to the relationships between 6 pairs of genes (see Table 1). However, no loop was observed between any pair of clusters. Here, a loop means that Gene 1 of cluster A regulates the expression of

Table 1. The relationships between genes for the regulation of expression

Gene A	Cluster no.	Gene B	Cluster no.		References
SNF2	9	SUC2	23	○	Neigeborn and Carlson (1984) <i>Genetics</i> , 108 , 845–858
SNF5	26	SUC2	23	○	Neigeborn and Carlson (1984) <i>Genetics</i> , 108 , 845–858
SNF6	5	SUC2	23	○	Neigeborn and Carlson (1984) <i>Genetics</i> , 108 , 845–858
SWI1	9	SUC2	23	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SWI3	9	SUC2	23	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SWI1	9	ADH1	30	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SNF2	9	ADH1	30	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SWI3	9	ADH1	30	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SWI1	9	ADH2	30	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SNF2	9	ADH2	30	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SWI3	9	ADH2	30	○	Peterson and Herskowitz (1992) <i>Cell</i> , 68 , 573–583
SNF2	9	HIS4	31	○	Jiang and Stillman (1995) <i>Genetics</i> , 140 , 103–114
GAL11	26	GAL2	19	×	Suzuki <i>et al.</i> (1988) <i>Mol. Cell. Biol.</i> , 8 , 4991–4999
GAL11	26	SUC2	23	○	Vallier and Carlson (1991) 129 , 675–684
GAL11	26	HIS3	14	×	Sakurai <i>et al.</i> (1996) <i>FEBS Lett.</i> , 398 , 113–119
GAL11	26	HIS4	31	○	Sakurai <i>et al.</i> (1996) <i>FEBS Lett.</i> , 398 , 113–119
SIN4	24	IME1	9	○	Shimizu <i>et al.</i> (1998) <i>Nucleic Acids Res.</i> , 26 , 2329–2336
SIN4	24	CTS1	27	○	Jiang <i>et al.</i> (1995) <i>Genetics</i> , 140 , 47–54
SIN4	24	HIS4	31	○	Jiang and Stillman (1995) <i>Genetics</i> , 140 , 103–114
SIN4	24	SUC2	23	○	Song and Carlson (1998) <i>EMBO J.</i> , 17 , 5757–5765
SFL1	19	SUC2	23	○	Song and Carlson (1998) <i>EMBO J.</i> , 17 , 5757–5765
SRB8	26	SUC2	23	○	Song and Carlson (1998) <i>EMBO J.</i> , 17 , 5757–5765
ROX3	20	SUC2	23	×	Song and Carlson (1998) <i>EMBO J.</i> , 17 , 5757–5765
RGR1	28	SUC2	23	○	Sakai <i>et al.</i> (1998) <i>Genetics</i> , 119 , 499–506
RGR1	28	CTS1	27	○	Jiang <i>et al.</i> (1995) <i>Genetics</i> , 140 , 47–54
SIN3	2	RME1	15	○	Vidal <i>et al.</i> (1991) <i>Mol. Cell. Biol.</i> , 11 , 6306–6316
SIN3	2	FUS1	18	○	Vidal <i>et al.</i> (1991) <i>Mol. Cell. Biol.</i> , 11 , 6306–6316
SIN3	2	BAR1	26	○	Vidal <i>et al.</i> (1991) <i>Mol. Cell. Biol.</i> , 11 , 6306–6316
SIN3	2	TRK2	2	–	Vidal <i>et al.</i> (1991) <i>Mol. Cell. Biol.</i> , 11 , 6306–6316
SIN3	2	PHO5	14	×	Vidal <i>et al.</i> (1991) <i>Mol. Cell. Biol.</i> , 11 , 6306–6316
MIG1	23	SUC2	23	–	Trumbly (1992) <i>Mol. Microbiol.</i> , 6 , 15–21
CYC8	2	SUC2	23	×	Trumbly (1992) <i>Mol. Microbiol.</i> , 6 , 15–21
TUP1	26	SUC2	23	○	Trumbly (1992) <i>Mol. Microbiol.</i> , 6 , 15–21
TUP1	26	RME1	15	○	Mukai <i>et al.</i> (1991) <i>Mol. Cell. Biol.</i> , 11 , 3773–3779
TUP1	26	ANB1	24	○	Zhang <i>et al.</i> (1991) <i>Gene</i> , 97 , 153–161
TUP1	26	ROX1	26	–	Zhang <i>et al.</i> (1991) <i>Gene</i> , 97 , 153–161
TUP1	26	CYC1	5	×	Zhang <i>et al.</i> (1991) <i>Gene</i> , 97 , 153–161
TUP1	26	IME1	9	○	Mizuno <i>et al.</i> (1998) <i>Curr. Genet.</i> , 33 , 239–247
RGR1	28	IME1	9	×	Shimizu <i>et al.</i> (1998) <i>Nucleic Acids Res.</i> , 26 , 2329–2336
RME1	15	IME1	9	×	Shimizu <i>et al.</i> (1998) <i>Nucleic Acids Res.</i> , 26 , 2329–2336

The experimental data for the regulatory relationship between a pair of genes were collected for the examination of our study with the currently available data. 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 numbers, to which Genes A and B belong, respectively. The fifth column includes three symbols, '○', '×', and '–'. A non-zero partial correlation coefficient between the corresponding clusters is regarded as agreeing with the experimental result, and '○' is put in the column. On the other hand, a zero partial correlation coefficient between the corresponding clusters is regarded as being inconsistent with the experimental result, and '×' is placed in the column. '–' in the fifth column indicates that both Genes A and B belong to the same cluster. To save space, the references for the experimental studies are written in the sixth column.

Gene 2 of cluster B, while Gene 3 of cluster B regulates the expression of Gene 4 of cluster A. We would like to discuss cluster 16 as another example for the introduction of arrows here, although cluster 16 is not included in Table 1 or Figure 3. The node corresponding to cluster 16 had 19 edges (see Figure 2). We do not have any experimental results about the regulatory relationships of the genes included in cluster 16. However, we could replace

all of the edges connected with cluster 16 with arrows toward the cluster, because cluster 16 did not include any genes that are classified as being involved in transcription by the *S. cerevisiae* Functional Catalogues of MIPS (<http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat>). All of the other 33 clusters included genes related to transcription, although the number of such genes was different from cluster to cluster.

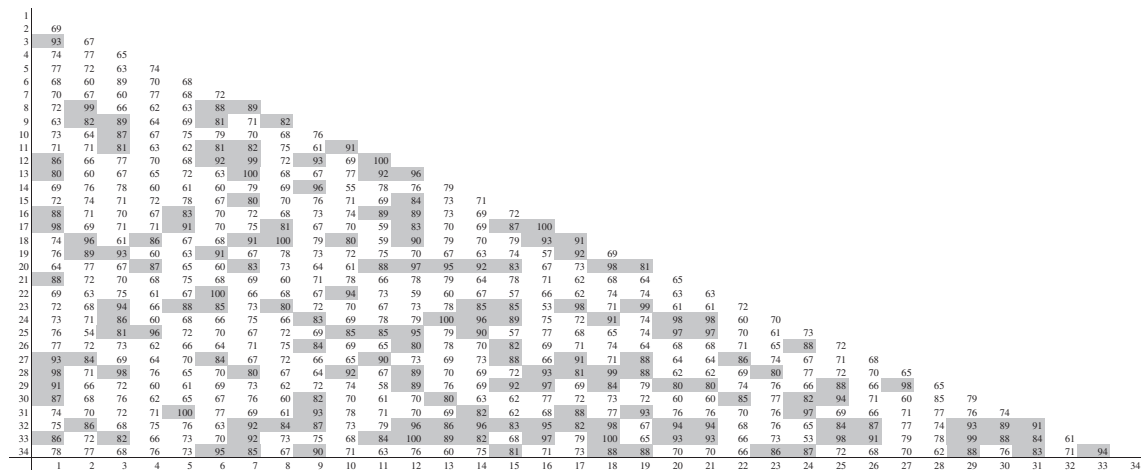


Fig. 4. Bootstrap probabilities of the edges. The result of the bootstrap test is shown, where the number of each element indicates the bootstrap probability for an edge appearance between a pair of clusters corresponding to the element. Each bootstrap probability is multiplied by 100 for percentage representation. When the probability is equal to or more than 80%, the element is shaded.

Evaluation of the inferred genetic network from statistical viewpoints

To evaluate the reliability of the edges of the obtained independence graph, or the non-zero elements of the obtained PCCM, the averaged profile data were subjected to a bootstrap analysis (Efron and Gong, 1982). Consider that the original sample was a data set of averaged expression levels of M clusters measured under N different conditions. In other words, the original sample consisted of the N sets of the averaged expression levels of the M clusters measured under the same condition. Then, a bootstrap sample was generated by randomly sampling N times, with replacement, from the original sample. The bootstrap sample was subjected to the analysis by GGM, and a PCCM for the bootstrap sample was obtained. This procedure was repeated K times, and we had K PCCMs for the bootstrap samples. Let's consider an element (i, j) of the original PCCM. It does not matter whether the element has a zero or non-zero value in the original PCCM. Then, the count of the non-zero values at element (i, j) over the K PCCMs for bootstrap samples was obtained. The ratio of the count against K is the bootstrap probability of the edge, or the reliability for the existence of the edge, of the element (i, j) . According to the same procedure, the bootstrap probability was obtained for each element. Here, K was set to 100.

The bootstrap probabilities thus obtained are shown in Figure 4. Here, we used 80% as the significance level for the bootstrap probability. Out of 561 elements, 173 elements had bootstrap probabilities $\geq 80\%$. Of the 173 elements, 163 corresponded to the non-zero elements of PCCM. The ratio was about 94%. That is, most of the

elements with the high bootstrap probability corresponded to the non-zero elements in the original PCCM. On the other hand, the original PCCM included 373 non-zero elements, and 163 out of them had bootstrap probabilities $\geq 80\%$. In other words, the edges in the independence graph, which corresponded to about 44% of the non-zero elements, were regarded as being statistically significant in this case.

As discussed above, there are many experimental results for the expression regulation of *S. cerevisiae*, and some edges in the obtained independence graph agreed with the experimental studies. However, even such edges sometimes had low bootstrap probabilities (see Figure 3). To check this problem further, the distributions of the bootstrap probability of the edge appearance for the zero and non-zero elements of the original PCCM were examined (see Figure 5). Clearly, the two distributions are different from each other. The averages of the bootstrap probability over the two distributions were 67.84 and 79.55, respectively. Welch's test suggested that the difference of the average was statistically significant, with a probability value of less than 0.001. However, the distribution of the non-zero elements had a long tail toward the lower value. In addition, relatively high bootstrap probabilities were sometimes observed even at the zero elements in the original PCCM. Therefore, the two distributions overlapped each other, and the overlap between the two distributions made it difficult to distinguish many edges from the non-adjacent relationships. As described above, the bootstrap probability is a good measure for the statistical significance of the edge in the independence graph. However, the situation described above suggests that it

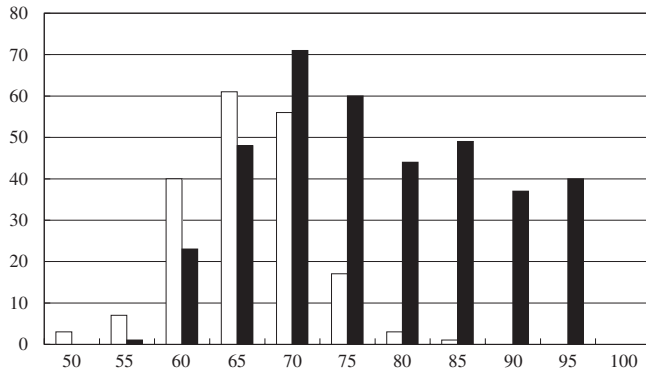


Fig. 5. The histogram of bootstrap probabilities for zero and non-zero elements of the original PCCM. As described in text, the original PCCM had 188 zero elements and 373 non-zero elements (see Figure 2). For each element, the bootstrap probability for edge appearance between a pair of clusters corresponding to the element was calculated (see Figure 4). The frequency distributions of the bootstrap probabilities for zero and non-zero elements are shown as a histogram. As shown in Figure 4, all of the obtained bootstrap probabilities were greater than 50%. Starting from the region of 50–55%, therefore, the frequency of zero elements in the original PCCM, with bootstrap probabilities that fell within each region of 5% in size, was counted. The same procedure was applied to the non-zero elements. The vertical axis indicates the frequency of the elements within the region, while the horizontal axis indicates the bootstrap probability. The open and filled bars indicate the frequency distributions for zero and non-zero elements, respectively.

is difficult to reconstruct genetic networks, when relying only on the edges with high bootstrap probabilities. The development of a statistical test, which has more power to distinguish pairs of conditionally independent clusters from those of conditionally dependent clusters, would be a future improvement of our approach.

There are other possible ways to represent the expression behaviors of the genes in the same cluster. For example, we randomly selected a representative gene from each cluster, and calculated the correlation coefficient matrix for the set of representative genes. Then, the PCCM of the 34 representative genes was calculated from the correlation coefficient matrix. The procedure was repeated 100 times, and a matrix averaged over the obtained 100 PCCMs was obtained. As described in the associated paper (Horimoto and Toh, 2001), most of the 100 correlation coefficient matrices thus generated could not be distinguished with statistical significance. However, the standard deviations of the elements of the averaged PCCM were very large. Therefore, we adopted the method described above. However, more effective ways to represent the expression behaviors of the clusters for the analysis by GGM would be possible, and the development of such methods would improve our method in the future.

Comparison with Bayesian networks

Recently, Friedman *et al.* (2000) reported the inference of a genetic network of *S. cerevisiae* with a method called Bayesian networks. Both GGM and Bayesian networks belong to a statistical analysis class called the ‘graphical model.’ They share some ideas such as the conditional independence and the Markov property, although the computational methods are quite different. Therefore, it is interesting to compare the results of the two approaches. In their paper, they used the term ‘Markov relation’ to represent the conditionally independent relationship between a pair of genes. Their Markov relation seems to correspond to a non-zero element of PCCM or an edge in the independence graph in our approach. Therefore, we compared the Markov relation among the *S. cerevisiae* genes shown in Tables 2 and 3 in the paper by Friedman *et al.* (2000) with the GGM results.

Both tables include the results of the Bayesian network estimation with the expression profile data by Spellman *et al.* (1998), although different assumptions are set for the estimations (see Friedman *et al.*, 2000 for details). In Table 2 of the paper by Friedman *et al.* (2000), 17 gene pairs are listed in the Markov relation category. Among them, one or both gene(s) of nine pairs were not included in the data we used. The genes of five pairs are present in the same clusters that we obtained. Only three pairs had genes that belong to different clusters. The genes that corresponded to two out of the three pairs were included in the clusters that had non-zero partial correlation coefficients, while the remaining one pair was estimated to be conditionally independent in our analysis. The situation was almost the same, even when Table 3 of the paper by Friedman *et al.* (2000) was used for comparison. Table 3 included 16 pairs of genes in the Markov relation category. Of these, 10 pairs could not be used for comparison, because one or both gene(s) of the pairs were not present in the data we used. In addition, 4 pairs also could not be used for comparison, because both genes of the pairs were included in the same clusters. The genes of the remaining 2 pairs were found in the different clusters. In our approach, the genes in one of the pairs had an edge in the independence graph, while the genes in another pair were estimated to be conditionally independent. Due to the small amount of data available for comparison, it was difficult to examine the correspondence between our result and the data by Friedman *et al.* As described above, we used the expression profile data measured by Eisen *et al.* (1998), while Friedman *et al.* (2000) used the expression profiles by Spellman *et al.* (1998). The difference in the data could be one of the reasons for the bad correspondence. On the other hand, the adoption of the cluster analysis in our approach seemed to decrease the resolution of the network description.

That is, the Bayesian network sometimes inferred the relationships between gene pairs, which were classified into the same clusters in our approach. Despite the low resolution, however, our approach could correctly infer the network of expression regulation to some extent, as discussed above. The improvement of the cluster analysis, by using better methodology or more detailed profile data, would increase the power of the network description of our approach. At this point, our approach would be useful as a preprocessing step for more detailed inferences of genetic networks.

Interpretation of edges in the independence graph

Generally speaking, a non-zero element of the PCCM just indicates the direct interaction between a pair of variables, but it cannot tell us which variable corresponds to cause or effect. Therefore, an edge corresponding to such an element in the independence graph is undirected. Without *a priori* knowledge about the causality, the edge cannot be replaced with an arrow. Therefore, a non-zero element is only considered to indicate the presence of a direct interaction between a pair of clusters. Then, what is the meaning of such an edge between clusters? As discussed above, the edges reflected the regulatory relationships between genes. However, it did not seem realistic that all of the genes included in a cluster affect the expression of all of the genes within the other cluster connected by an edge. It would be more plausible that only a subset of the genes in one cluster directly affects the expression of a subset of genes in the other cluster. Let's consider two clusters, A and B, connected by an edge. Then, a subset of genes within cluster A may affect the expression of a subset of genes within cluster B.

In addition, we should keep in mind that the expression profiles of many function-unknown genes are not included in the data by Eisen *et al.* (1998). In the work by Eisen *et al.* (1998), the expression levels of the 2467 genes of *S. cerevisiae* whose functions have been characterized are measured. However, there are 6241 genes predicted to encode proteins in the yeast genome (Rubin *et al.*, 2000). Due to the missing data, we could not exclude the possibility that some edges may be generated for gene pairs that interact indirectly, through the expression of the function-unknown genes.

Concluding remarks

In this manuscript, we report the combined application of cluster analysis and GGM to infer genetic networks from expression profile data. The final goal of the inference of the genetic network is the complete description of the causality of the expression of all of the genes in a genome, that is, the inference of the full relationships between the transcription-related genes and all of the genes in a genome. Our approach with the expression profile data

available today did not attain an inference of such high resolution. We were only able to infer the relationship among clusters of genes. However, our study suggested that even such a low resolution inference can explain an experimental study of transcriptional regulation to some extent, although the improvement of the resolution is a goal in the future. Several assumptions have been introduced for the application of GGM. For example, the expression profile data are assumed to be drawn from a multivariate normal distribution. Such assumptions should be re-examined to improve the resolution of the inference.

Finally, we would like to conclude this manuscript with a future extension of the current approach. As described above, one of the important problems in studying expression profile data is the inference of causality in the genetic network. In order to introduce the causality into the independence graph, some information other than the expression profile is required. In this manuscript, the edges of the subgraph were replaced with arrows, according to the previous experimental results. When time series data are provided for GGM, however, we can systematically introduce the causality according to the time order. The graph obtained by this approach is called a 'chain independence graph.' However, some modifications of the GGM algorithm are required for the inference of the genetic network as a chain independence graph. In addition, the subjects of GGM application are not restricted to the expression profile data. For example, GGM could be applicable to the inference of the contact sites from a multiple alignment. GGM has high potential for investigations of interactions in the field of bioinformatics.

ACKNOWLEDGEMENTS

One of the authors (K.H.) was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas (C) 'Genome Information Science' from the Ministry of Education, Science, Sports, and Culture of Japan (grant 12208038).

REFERENCES

- Akutsu,T., Miyano,S. and Kuhara,S. (2000) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comput. Biol.*, **7**, 331–343.
- Ben-Dor,A., Shamir,R. and Yakhini,Z. (1999) Clustering gene expression patterns. *J. Comput. Biol.*, **6**, 281–297.
- Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Proc. Pac. Symp. Biocomput.*, 17–28.
- D'haeseleer,P., Wen,X., Fuhrman,S. and Somogyi,R. (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Proc. Pac. Symp. Biocomput.*, 41–52.
- Edwards,D. (1995) *Introduction to Graphical Modelling*. Springer, New York.

- Eisen,M.B., Spellman,P.T., Prown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.
- Efron,B. and Gong,G. (1982) A leisurely look at the bootstrap, the jackknife and cross-validation. *Am. Stat.*, **37**, 36–48.
- Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
- Horimoto,K. and Toh,H. (2001) Statistical estimation of cluster boundaries in gene expression profile data. *Bioinformatics*, in press.
- Li,Y., Bjorklund,S., Jiang,Y.W., Kim,Y.J., Lane,W.S., Stillman,D.J. and Kornberg,R.D. (1995) Yeast global transcriptional regulators Sin4 and Rgr1 are components of mediator complex/RNA polymerase II holoenzyme. *Proc. Natl Acad. Sci. USA*, **92**, 10 864–10 868.
- Peterson,C.L. and Herskowitz,I. (1992) Characterization of the yeast SWI1, SEI2, and SWI3 genes, which encode a global activator of transcription. *Cell*, **68**, 573–583.
- Rubin,G.M., Yandell,M.D., Wortman,J.R., Gabor Miklos,G.L., Nelson,C.R., Hariharan,I.K., Fortini,M.E., Li,P.W., Apweiler,R., Fleischmann,W., Cherry,J.M., Henikoff,S., Skupski,M.P., Misra,S., Ashburner,M., Birney,E., Boguski,M.S., Brody,T., Brokstein,P., Celniker,S.E., Chervitz,S.A., Coates,D., Cravchik,A., Gabrielian,A., Galle,R.F., Gelbart,W.M., George,R.A., Goldstein,L.S., Gong,F., Guan,P., Harris,N.L., Hay,B.A., Hoskins,R.A., Li,J., Li,Z., Hynes,R.O., Jones,S.J., Kuehl,P.M., Lemaitre,B., Littleton,J.T., Morri-son,D.K., Mungall,C., O'Farrell,P.H., Pickeral,O.K., Shue,C., Vossball,L.B., Zhang,J., Zhao,Q., Zheng,X.H., Zhong,F., Zhong,W., Gibbs,R., Venter,J.C., Adams,M.D. and Lewis,S. (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.
- Somogyi,R. and Shiegoski,C.A. (1996) Modeling the complexity of genetic networks: understanding multigene and pleiotropic regulation. *Complexity*, **1**, 45–63.
- Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression pattern with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
- Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
- Tzamarias,D. and Struhl,K. (1994) Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex. *Nature*, **369**, 758–761.
- Wermuth,N. and Scheidt,E. (1977) Fitting a covariance selection to a matrix. Algorithm AS 105. *Appl. Stat.*, **26**, 88–92.
- Whittaker,J. (1990) *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.