
Journal of Biological Physics **28:** 449–464, 2002.

System for Automatically Inferring a Genetic Network from Expression Profiles

H. TOH[1] and K. HORIMOTO[2]
[1] *Department of Bioinformatics, Biomolecular Engineering Research Institute 6-2-3 Furuedai, Suita, Osaka 565-0874, Japan*
[2] *Laboratory of Mathematics, Saga Medical School, 5-1-1 Nabeshima, Saga, Saga 849-8501, Japan*

Abstract. A system is constructed to automatically infer a genetic network by application of graphical Gaussian modeling to the expression profile data. Our system is composed of two parts: one part is automatic determination of cluster boundaries of profiles in hierarchical clustering, and another part is inference of a genetic network by application of graphical Gaussian modeling to the clustered profiles. Since thousands of or tens of thousands of gene expression profiles are measured under only one hundred conditions, the profiles naturally show some similar patterns. Therefore, a preprocessing for systematically clustering the profiles is prerequisite to infer the relationship between the genes. For this purpose, a method for automatic determination of cluster boundaries is newly developed without any biological knowledge and any additional analyses. Then, the profiles for each cluster are analyzed by graphical Gaussian modeling to infer the relationship between the clusters. Thus, our system automatically provides a graph between clusters only by input the profile data. The performance of the present system is validated by 2467 profiles from yeast genes. The clusters and the genetic network obtained by our system are discussed in terms of the gene function and the known regulatory relationship between genes.

Key words: cluster analysis, cluster boundary, gene expression profile, genetic network, graphical Gaussian modeling, microarray

1. Introduction

Advances in microarray techniques have enabled us to measure whole-genome mRNA abundance [1–4]. The expression levels of thousands or tens of thousands of genes can be simultaneously monitored under multiple conditions. These gene expression profile data are compiled at several databases, and are available at their web sites. This provides an enormous opportunity to elucidate the underlying information in the complex data for functional genomics and proteomics.

An essential step in the analysis of gene expression profile data is the detection of gene groups that manifest similar expression patterns. Several techniques have been used for detecting similar expression patterns [5–10]. Hierarchical clustering is clearly valuable. One of the merits of hierarchical clustering is the visual presentation that enables us to intuitively understand the clustering of genes in a dendrogram, where some genes that are mutually related in terms of the cell

function are grouped into the same cluster [5, 11]. Indeed, the cluster boundaries for the interpretation of the profile patterns are determined by visual inspection of the dendrogram. However, there are some dendrograms where the nodes are connected by very short branches, due to the highly correlated gene expression profiles. Subsequently, the cluster boundaries are determined with the help of some exploratory methods, such as biological knowledge of the genes and sequence analyses of the upstream regions of genes [5–7, 11, 12]. An alternative to detect the similar patterns is a K-means clustering algorithm [13, 14]. In the algorithm, the cluster boundaries are determined by the optimization of some statistical criteria, such as the maximum variance of clusters, without the manual intervention and the arbitrary thresholds. However, a given number of clusters, K, is prerequisite to analyze the samples in the algorithm, and therefore the determination of the cluster boundaries requires repetitive clustering for many different K values [10, 12]. Thus, it remains a challenge to systematically estimate the cluster boundaries in the clustering.

Another step in the profile analysis is the inference of the regulatory networks among genes, which here is called the 'genetic network'. Modelings with the Boolean network [15], differential equations [16, 17], and a combination of the methods [18] have been investigated for inferences of the genetic networks. An approach, which combines cluster analysis with sequence motif detection, to determine the genetic network architecture is also proposed [10]. Recently, an approach to infer the genetic networks with Bayesian networks was proposed [19].

Recently, we have proposed a novel approach to infer the genetic networks from the expression profiles [20] in the combination with a newly developed method for the automatic clustering [21]. In our approach, the genetic networks are inferred by a combination of cluster analysis and a method called 'Graphical Gaussian Modeling' (GGM) [22, 23]. Here, two methods are synthesized to a system for automatically inferring a genetic network only by the input of the expression profile data. The validity of the system is discussed from both biological and statistical viewpoints.

2. Materials and Methods

2.1. EXPRESSION PROFILE DATA

The gene expression profile data analyzed here are cited from Eisen *et al.* [5] (http://www.pnas.org or http://rana.stanford.edu/clustering/). The data comprise the expression profiles of 2467 yeast (*Saccharomyces cerevisiae*) genes that were measured under 79 conditions.

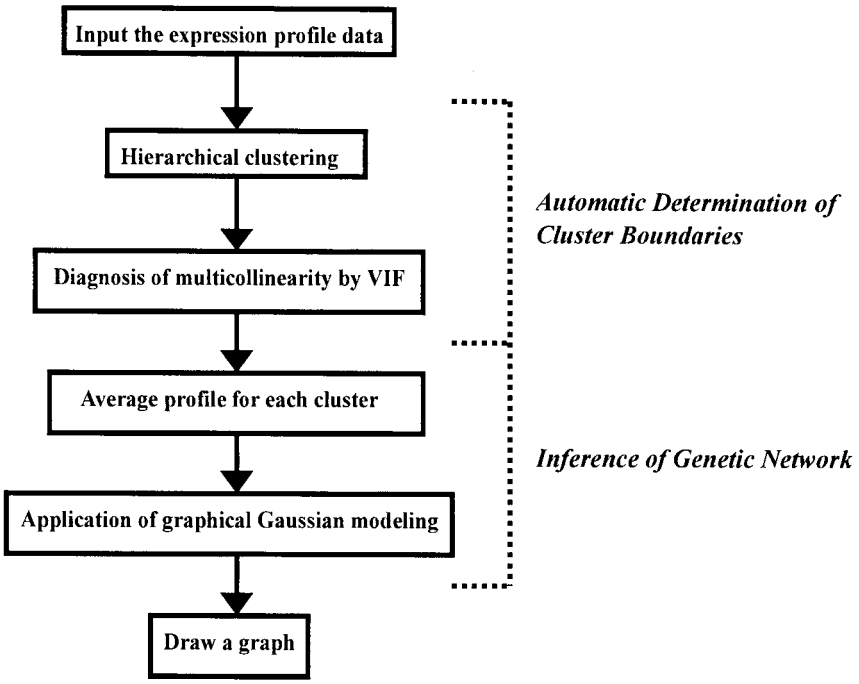

Figure 1. Flow of the Algorithm for the Present System. The details of each step are described in Materials and Methods.

2.2. PROCEDURE OF THE PRESENT SYSTEM

The present procedure is composed of two parts: automatic determination of cluster boundaries and the systematic inference of relationship between the clusters. The procedure is depicted in Figure 1. The details of the two parts are described below.

Algorithm for determining cluster boundaries

The present algorithm is composed of five steps in two parts. Each step is described below.

Step 1: For clustering the profile data, a metric is defined to measure the similarity between the expression profiles. In the present analysis, the Euclidean distance between the Pearson correlation coefficients is adopted as the metric, i.e.,

$$d_{ij} = \sqrt{\sum_{l=1}^{n} (r_{il} - r_{jl})^2} \tag{1}$$

where n is total number of the genes, and r_{ij} is the Pearson correlation coefficient between the i and j genes of the expression profile that are measured at m points, p_{ik}, $(k = 1, 2, \ldots, m)$:

$$r_{ij} = \frac{\sum_{k=1}^{m}(p_{ik} - \overline{p_i})(p_{jk} - \overline{p_j})}{\sqrt{\sum_{k=1}^{m}(p_{ik} - \overline{p_i})^2 \cdot \sum_{k=1}^{m}(p_{jk} - \overline{p_j})^2}} \tag{2}$$

where $\overline{p_i}$ is the arithmetic average of p_{ik} over m points. The above distance is used to evaluate the similarity between genes in terms of the expression pattern. The smaller the distance is between the two genes, the more similar the corresponding genes are in the expression profile patterns. Notably, the present distance between the two genes is designed to reflect the similarity in the expression profile patterns between other genes as well as between the measured points. The distances defined in the Equation (1) are analyzed by a standard hierarchical clustering technique, the group average method [14, 24]. By the group average clustering, $(n - 1)$ dissimilarity scores of the nodes are obtained, i.e.,

$$\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_{n-1}, \quad \text{where } \hat{d}_i < \hat{d}_j, \quad \text{if } i < j.$$

Step 2: A correlation coefficient matrix is generated from the original correlation coefficient matrix, at each node in the hierarchical clustering. For example, when $\hat{d}_c$ is set to be $\hat{d}_{2467-m+1} \geqq \hat{d}_c > \hat{d}_{2467-m}$, m clusters are obtained at the $(2467 - m)$ node. In the m clusters, the members within each cluster share a simlilar expression pattern by the clustering procedure. When the genes clustering procedure. When the genes are numbered from 1 to 2467 in the original profile data, therefore, the gene with the youngest number is selected among the members of a cluster. Thus, an $m \times m$ correlation coefficient matrix is obtained at the $(2467 - m)$ for m clusters. The robustness of the selection procedure is discussed in a subsequent section.

Step 3: A statistical property of the $m \times m$ correlation coefficient matrix is evaluated at the $(2467 - m)$ node in the dendrogram obtained in Step 2, with the use of the variance inflation factor (VIF) in the multiple regression analysis. In the multiple regression analysis, a criterion variable is generally expressed by a linear combination of multiple explanatory variables, i.e.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n, \tag{3}$$

where y and x_i are criterion and explanatory variables, respectively, and β_i is the regression coefficient of the corresponding variable. One result of using a large number of explanatory variables is that many of the variables are highbly correlated. The existence of high correlations among the explanatory

valuables is known as multicollinearity, and the variables that are involved in the multicollinearity are diagnosed by VIF, as follows,

$$\mathrm{VIF}_i = r_{ii}^{-1}, \tag{4}$$

where r_{ii}^{-1} is the ith diagonal element of the inverse matrix of correlation coefficients between explanatory variables [25]. In a correlation coefficient matrix between m explanatory variables, therefore, m VIF's are calculated.

Step 4: The VIF is applied to estimate the cluster boundaries in the expression profile data. When the explanatory variables in the Equation (3) correspond to the gene profiles, the VIF expresses the degree of linear relationship between the profiles. In the diagnosis of the multicollinearity, the popular cutoff value of 10.0 [25, 26] is adopted as a threshold in the present analysis: when VIF_i is larger than 10.0, the multicollinearity of the ith variable exists. Although the criterion variable, y in Equation (3), is a hypothetical variable in the present profile data, this is not an obstacle to the practical calculation of the VIF that is obtained from the inverse matrix of correlation coefficients between the profiles. Instead of the popular cutoff value of 10.0, $1/(1 - R_i^2)$ can also be set as a threshold for the multicollinearity, where R_i^2 is the coefficient of determination of the regression of x_i, on all other explanatory variables in the Equation (3). However, the practical profile data for the criterion variable are needed for the calculation of R_i^2. When designed profile data are obtained by suspension of the expression of a selected gene, the profiles of the expression-suspended gene can be set as a criterion variable. In this case, $1/(1 - R_i^2)$ may be adopted as a threshold for the multicollinearity.

Step 5: The m VIF's are assessed with the following condition:

$$\max\{\mathrm{VIF}_i\} < 10.0 \ \text{ for } \ i = 1, 2, \ldots, m.$$

If the condition is satisfied, then no multicollinearity exists in the m sets of profiles. In contrast, if the condition is not satisfied, then multicollinearity still exists in the profiles.

The above Steps from 2 to 5 proceed in an ascending order of nodes from 1 to 2466, and the first node that satisfies the above condition is searched. Thus, the maximum number of clusters is obtained. In other words, the maximum cluster number is searched, where each set of profiles shows no linear relationship. In the clusters, the genes are clearly classified into each cluster so that the cluster boundaries are determined. Notably, the cluster boundary determinations depends on only the properties of the profile data, without any additional analyses in the hierarchical clustering and various initializations for each K value in the K-means algorithm.

Conceptual framework of graphical Gaussian modeling

The correlation coefficient in the expression profile data has been widely utilized to evaluate the distance between genes for the cluster analysis [5]. Suppose that a pair of genes, say Genes A and B, show high correlation in their expression profiles. There are three possible mechanisms to induce 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 the third type of interactions would also cause 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 [22, 23], 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 correctly speaking, an edge in the independence graph indicates a pair of variables that are conditionally dependent.

Algorithm for application of graphical Gaussian modeling

To infer the relationship among variables from a set of observed data of the variables, at first, the calculaflon of the inverse of the covariance matrix Σ is required for GGM. However, many genes share similar expression patterns [5], and a high similarity in expression pattern induces linear dependence among rows or columns in the correlation coefficient matrix, in terms of numerical analysis. This causes to make the calculation of the inverse matrix difficult. Instead of raw expression profile data of genes, therefore, the averaged expression levels of clusters are hereafter considered. 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 (p(cluster 1(i)), p(cluster 2(i)), ..., p(cluster M(i))) and $1 \leq i \leq N$. The averaged expression level of the cluster k at the j-th condition is calculated as follows;

$$p(\text{cluster } k(j)) = (\Sigma_{\text{gene } i(j)\in\text{cluster } k} (p(\text{gene } i(j)))/n_k \qquad (1 \leq k \leq M,\ 1 \leq j \leq N)$$

where n_k is the number o members of the cluster k.

The averaged profile data was subjected to the analysis by GGM.

In actual application of GGM, we applied a stepwise and iterative algorithm developed by Wermuth and Scheidt [27], in order to evaluate which pair of clusters is conditionally independent.

Step 0: A complete graph of G(0) = (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 whose averaged expression levels are conditionally dependent, given the rest. 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 $P(\tau)$ from the correlation coefficient matrix $C(\tau)$. τ indicates the number of the iteration. The elements of the covariance matrix are calculated with the elements of the inverse of the original $M \times M$ covariance matrix. Let Ω (ω^{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^{ij}, and the off-diagonal element is given as ω^{ij}. If $\omega^{ij} = 0$, the conditional normal distribution is expressed as a product of the function of the averaged expression level of cluster i and that of cluster j. That is, clusters i and j is conditionally independent in expression level, 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). $\rho^{ij,\text{ the rest}}$ is given as $-\omega^{ij}/(\omega^{ii} \times \omega^{jj})$. That is, the variables i and j is conditionally independent when $\rho^{ij,\text{ the rest}} = 0$.

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 as the same as those of $C(\tau)$.

Step 4: In Wermuth and Sheidt algorithm, the termination of the iteration is judged by the values called 'deviance'. We here used two types of deviance, $dev1$ and $dev2$, whose definitions are as follows;

$$dev1 = N \log(|C(\tau + 1)| \,/\, |C(0)|),$$

$$dev2 = N \log(|C(\tau + 1)| \,/\, |C(\tau)|),$$

Calculate $dev1$ and $dev2$. 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 $dev1 \leq 0.05$, or the probability value corresponding to $dev2 \leq 0.05$, 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 $\mathrm{G}(\tau)$ to generate $\mathrm{G}(\tau + 1)$, and τ is increased by 1. Then, go to Step 1.

The graph obtained by the 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. That is, the graph represents the genetic network of the M clusters under consideration.

3. Results

3.1. DETERMINATION OF CLUSTER BOUNDARIES

Following the hierarchical clustering of 2467 profiles, the cluster boundaries are estimated by the calculation of VIF at each node in the dendrogram. The fraction of the number of VIF's that are less than 10.0 to the total number of diagonal elements in correlation coefficient matrix is plotted against the cluster number in Figure 2. In the calculation of VIF, the inverse of the correlation coefficient matrix cannot be calculated in more than 50 clusters (49 nodes). This is because some values in the process of implementation are too small to proceed with the calculation of an inverse matrix. Thus, the VIF is calculated in less than 49 clusters.

As seen in Figure 2, the fractions of the diagonal elements with VIF's less than 10.0 increase as the number of clusters decreases. This indicates that the multicollinearity between the profiles monotoneously diminishes with the decrease of clusters. Although the fractions of VIF's were not calculated in more than 50 clusters, due to the limitation of the present numerical analysis, the monotoneous decrease suggests that the fraction may show a small value in more than 50 clusters in the present analysis. Finally, the fraction reaches a 1 value in 34 clusters, and maintains a 1 value in fewer clusters. This indicates that less than 34 clusters show no linear relationships between the profile data. Consequently, the maximum number of clusters with all VIF's less than 10.0 is estimated to be 34.

In the 34 clusters, the members are ranged from 14 to 275. The list of all members, in addition to the dendrogram of 34 clusters, is available at our web sites

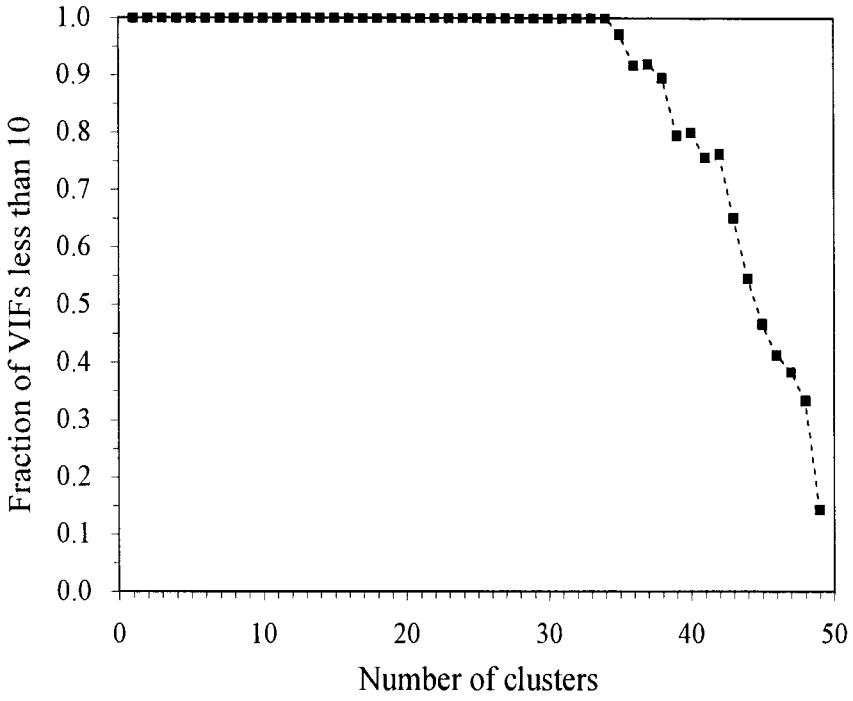

Figure 2. Estimation of Cluster Boundaries by the Variance Inflation Factor. The fractions of the diagonal elements with less than 10.0 of VIF to all diagonal elements with a number equal to the cluster number are plotted from 49 clusters to 2 clusters. In more than 50 clusters, the inverse of the correlation coefficient matrix cannot be numerically calculated, due to the highly correlated profiles of the genes.

(http//www.ged.sagamed.ac.jp/horimoto/microarray/ or http:www.beri.or.jp/toh/~protein).

3.2. APPLICATION OF GGM

The GGM was applied to the average expression levels of 34 clusters obtained in the preceding subsection. The iterative procedure of GGM was stopped when either probability value for $dev1$ or $dev2$ did not satisfy the given level of significance. In the present analysis, the probability values for $dev1$ were always greater than 99.999% over the iterative calculation. In contrast, those of $dev2$ gradually decreased with increasing step numbers. Finally, the iterative calculation was stopped at the step number $= 189$. The probability value for $dev2$ at the 189-th 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. Consequently, $dev2$ was effective for the judgment of stopping the iteration, while $dev1$ was not a good measure for the judgment.

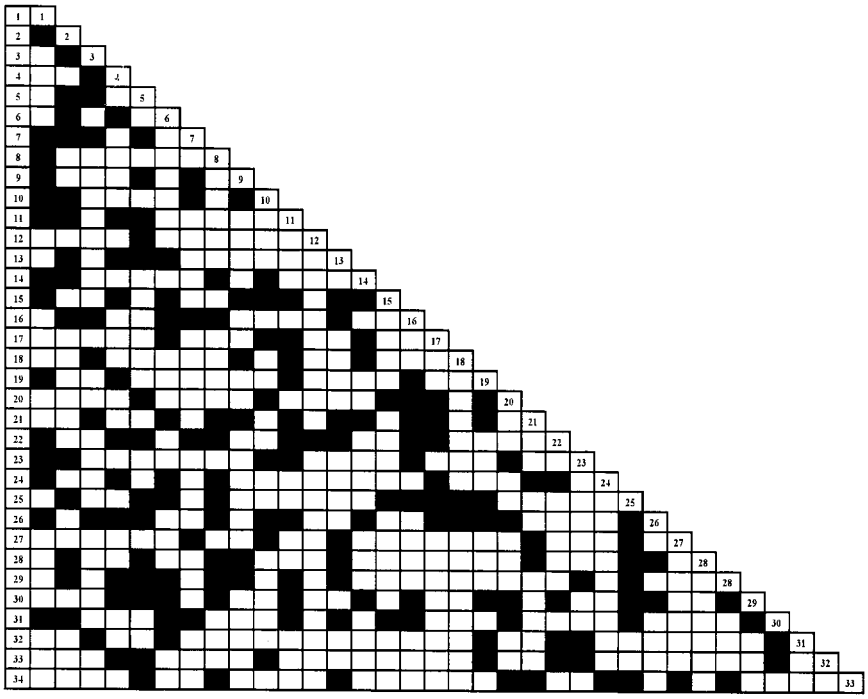

Figure 3. Schematic Presentation of Partial Correlation Coefficient Matrix Obtained by GGM. The partial correlation coefficient of every pair of 34 clusters is schematically shown, where the elements replaced with 0.0 in the iterative procedure of GGM are closed, and the others are open. The rows or columns correspond to the clusters, and the cluster numbers are shown at the left and diagonal of the matrix.

The results of application of GGM to the 34 clusters are shown in Figure 3. Since the expression levels are averaged in the same clusters, the values of partial correlation coefficients did not always reflect the degree of regulations experimentally observed. In Figure 3, therefore, the zero or non-zero features of the elements in the obtained PCCMs were focused. In other words, we will examine the accuracy of our approach, only based on presence or absence of edges in the independence graph. Hereafter, the presence of an edge in the independence graph is used as the same meaning as a non-zero partial correlation coefficient in PCCM.

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.

4. Discussion

4.1. EVALUATION OF CLUSTER BOUNDARIES

The cluster boundaries determined by the present analysis are evaluated by an investgation of each member of the clusters in terms of gene function. As described by Eisen *et al.* [5], nine groups of functionally related genes, defined according to the functional annotation in the *Saccharomyces* Genome Database [28], have been picked up by visual inspection of the entire clustered image, although the clustering boundaries were not clearly shown in their work. Most of the genes belonging to the nine groups in the previous paper are clustered together, and are allocated into the distinctive clusters in the present study. Indeed, 124 genes among 135 genes picked out by Eisen *et al.* [5] are correctly allocated. Furthermore, the present results were evaluated according to the gene classification scheme in the MIPS Yeast Genome Database [29]. Although the classification scheme in the MIPS Database is different from that in the *Saccharomyces* Genome Database, the cluster members in the present analysis are corresponded to the genes in the categories of the MIPS Database, which are similar to the gene groups picked up in Eisen *et al.* [5]. In six groups, the most frequent genes are consistently found in the clusters picked out by Eisen *et al.* [5]. Indeed, the most frequent genes are found in clusters 7, 10, 11, 12, 30 and 31, respectively. The two remaining categories are not consistent with the previous results. This is partly because the classification schemes are different between the two databases, partly because the genes in the two categories may be involved in multiple biological processes, and partly because the number of clusters may overestimate the underlying diversity of gene expression classes in the present data. At any rate, the automatically determined clusters in the present analysis agree well with the previous results that were obtained by visual inspection and biological knowledge of gene function. The complete correspondence of the cluster members with the categories in the MIPS scheme is available at our web sites.

The estimation of cluster boundaries is based on the correlation coefficient matrix whose elements are selected from the original correlation coefficient matrix at each node. To test the robustness of the present cluster boundaries, therefore, the correlation coefficient matrix used in the present study is compared with two types of matrices. One is the matrix that is generated by the random selection of correlation coefficients from the members of each cluster, except for the youngest number of genes used in the present study, and another is the matrix whose elements are the averaged correlation coefficients over the members in the respective clusters. The difference was very small between the correlation coefficient matrix in the present study and the randomly generated correlation coefficient matrix. The deviance with $P < 0.05$ is found in only 7 out of 100 comparisons. In contrast, the deviance with $P > 0.90$ is found in 82 comparisons. As for the matrix of the averaged correlation coefficients, there was also little difference from the matrix in the present study. The probability of the deviance between the two matrices is

0.985 ($\chi^2 = 489.424$). These results indicate that the correlation coefficient matrix generated by the present selection procedure appropriately represents the statistical properties in each cluster, and thus provides a vigorous estimation of the cluster boundaries.

4.2. EVALUATION OF THE INFERRED GENETIC NETWORK

Here, we examined the correspondence of the edges with the regulatory relationships directly or indirectly suggested by experimental studies. However, it was difficult to collect all of the literatures of experimental studies about the regulation of expression in *S. cerevisae*, because many experimental results have been accumulated. Thus, we collected the literatures, focusing on the regulation of SUC2 expression, because SUC2 is one of the genes that have been extensively studied. Some literatures about the expression of different genes from SUC2 were found during the collection process, which were also used for the examination of our result. Then, we evaluated obtained PCCM with the results of the collected experimental studies, under the assumption that the relationships obtained by experiments reflect the direct interactions about the expressions of the genes. When the partial correlation coefficient between two clusters, corresponding to a pair of the genes described in a 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.

A subgraph of the independence graph is shown in Figure 4. Each node corresponds with with a cluster, although only the genes related to SUC2 expression are written in the nodes. Both correct and wrong relationships are included in the subgraph. SUC2 is a gene for sucrose hydrolyzing enzyme called invertase, which is included in cluster 23. SNF1, 2 and 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 [30]. 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, 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 (another name is SPT13 or RAP3) has been identified as a transcription 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 (another name is BEL2) and RGR1 are considered to form a complex for transcription regulation [31]. SIN4 is included in cluster 24, while RGRI belongs to cluster 28. Both of them are involved in the regulation of SUC2 expression. The presence of edges between cluster 23 and clusters 24 and 28 agreed with the observation. TUP1, CYC8, and MIG1 are also considered to form a complex for regulation of glucose repression related genes [32]. TUP1 belongs to cluster 26. CYC8 (another name is SSN6) is included in

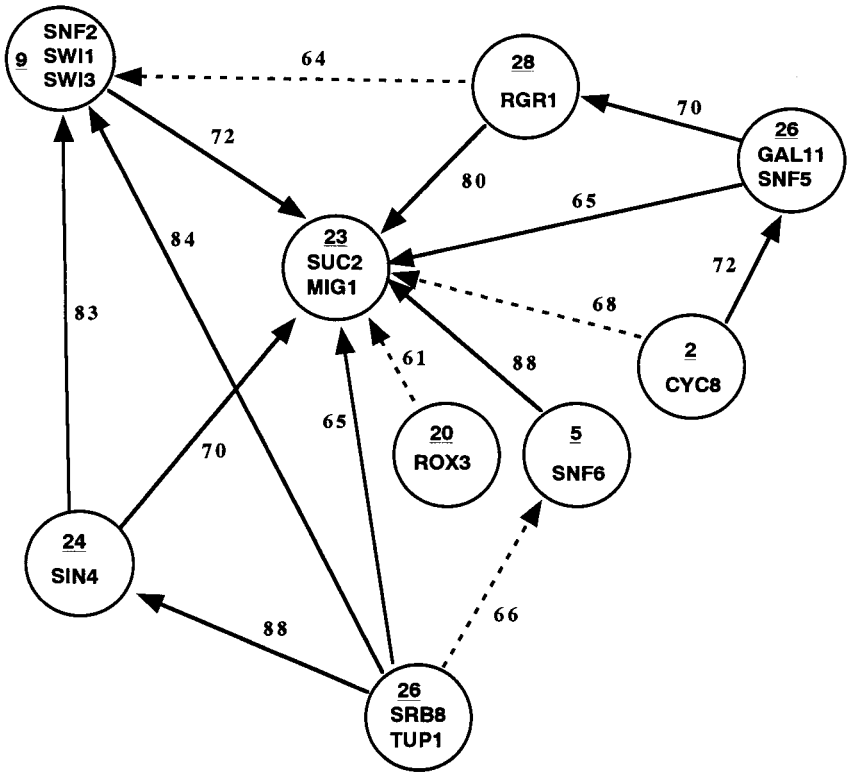

Figure 4. Correspondence between Inferred Network and Known Regulatory Relationship between the Genes Related to SUC2. A solid line indicates the interaction between a pair of clusters, which are also suggested by our approach. Each node indicates a cluster. A dashed line indicates the regulatory relationship among 34 clusters. 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. 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 the genes are involved in the regulation of SUC2 expression.

cluster 2. MIG1 is included in cluster 23, the same cluster as SUC2. The edge between cluster 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 3). SRB8 is involved in the SUC2 expression, which belongs to cluster 26, like TUP1. Thus, most of the collected experimental studies about the regulation of SUC2 are consistent with the result of GGM. Likewise, most of the remaining edges were consistent with the collected expression regulatory relationships other than those of SUC2.

The reliability of the edges of the obtained independence graph, or the non-zero elements of obtained PCCM is evaluated by the bootstrap analysis [33]. Consider

that the original sample was a data set of averaged expression levels of M clusters measured under N different conditions. 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. 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 above procedure, the bootstrap probability was obtained for each element. Here, K was set to 100. Given that 80% as the significance level for the bootstrap probability, out of 561 elements, 173 elements had bootstrap probabilities $\geq$ 80%. 163 out of the 173 elements 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 the 373 elements had the 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.

4.3. CONCLUDING REMARKS

We report a combined application of the 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 expressions of all of the genes in a genome, that is, the inference of the full relationships between 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 the experimental study for the transcription regulation to some extent, although improvement of the resolution is one of the important goals in the next step. 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, a future extension of the current approach is shortly discussed. 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 expression profile is required. In this work, 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 the approach is called a 'chain independence graph'. However, some modifications of the algorithm of the GGM are required for the inference of the genetic network as a chained 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. Thus, GGM has high potential for investigations of interactions in the field of bioinformatics.

Acknowledgement

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 13208028).

References

1. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L.: DNA expression monitoring by hybridization to high density oligonucleotide arrays, *Nature Biotechnol.* **14** (1996), 1657–1680.
2. Shalon, D., Smith, S.J. and Brown, P.O.: A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, *Genome Res.* **6** (1996), 639–645.
3. DeRisi, J., Iyer, V. and Brown, P.: Exploring the metabolic genetic control of gene expression on a genomic scale, *Science* **278** (1997), 680–686.
4. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *Sacchromyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell.* **9** (1998), 3273–3297.
5. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* **95** (1998), 14863–14868.
6. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R.: Large-scale temporal gene expression mapping of central nervous system development, *Proc. Natl. Acad. Sci. USA* **95** (1998), 334–339.
7. Alon, U., Barkai, N., Notterman, D.A., Gish, G., Ybarra, S., Mack, D. and Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA* **96** (1999), 6745–6750.
8. Ben-Dor, A., Shamir, R. and Yakhini, Z.: Clustering gene expression patterns, *J. Comput. Biol.* **6** (1999), 281–297.
9. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R.: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA* **96** (1999), 2907–2912.
10. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M.: Systematic determination of genetic network architecture, *Nature Genet.* **22** (1999), 281–285.
11. Eisen, M.B. and Brown, P.O.: DNA arrays for analysis of gene expression, *Methods Enzymol.* **303** (1999), 179–205.
12. Vilo, J., Brazma, A., Jonassen, I., Robinson, A. and Ukkonen, E.: Mining for putative regulatory elements in the yeast genome using gene expression data, in: R. Altman, T.L. Bailey, P. Bourne,

M. Gribskov, T. Lengauer, I.N. Shindyalov, L.F. Ten Eyck and H. Weissig (eds.), *Proceedings of Eighth International Conference on Intelligent Systems for molecular Biology*, AAAI Press, Menlo Park, 2000, pp. 384–394.
13. Hartigan, J.A.: *Clustering Algorithms*, Wiley, New York, 1975.
14. Gordon, A.D.: *Classification*, Chapman and Hall, London, 1981.
15. Somogyi, R. and Shiegoski, C.A.: Modeling the complexity of genetic networks: Understanding multigene and pleiotropic regulation, *Complexity* **1** (1996), 45–63.
16. Chen, T., He, H.L. and Church, G.M.: Modeling gene expression with differential equations, *Proc. Pacific Symp. Biocomput.* (1999), 17–28.
17. D'Haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R.: Linear modeling of mRNA expression levels during CNS development and injury, *Proc. Pacific Symp. Biocomput.* (1999), 41–52.
18. Akutsu, T., Miyano, S. and Kuhara, S.: Algorithms for inferring qualitative models of biological networks, *Proc. Pacific Symp. Biocomput.* (2000), 290–301.
19. Friedman, N., Linial, M., Nachman, I. and Pe'er, D.: Using Bayesian networks to analyze expression data, *J Comp. Biol.* **7** (2000), 601–620.
20. Toh, H. and Horimoto, K.: Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, in press.
21. Horimoto, K. and Toh, H.: Statistical estimation of cluster boundaries in gene expression profile data, *Bioinformatics*, **17** (2001), 1143–1151.
22. Whittaker, J.: *Graphical Models in Applied Multivariate Statistics*, John Wiley, Chichester, 1990.
23. Edwards, D.: *Introduction to Graphical Modelling Second Edition*, Springer, New York, 2000.
24. Sneath, P.H.A. and Sokal, R.R.: *Numerical Taxonomy*, W.H. Freeman and Company, San Francisco, 1973.
25. Freund, R.J. and Wilson, W.J.: *Regression Analysis*, Academic Press, San Diego, 1998.
26. Chatterjee, S. and Price, B.: *Regression Analysis by Examples*, John Wiley & Sons, New York, 1977.
27. Wermuth, N. and Scheidt, E.: Fitting a covariance selection to a matrix. Algorithm AS 105, *Appl. Statist.* **26** (1977), 88–92.
28. Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K., *et al.*: Genetic and physical maps of *Saccharomyces cerevisiae*, *Nature* **387** (1997), 67–73.
29. Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., *et al.*: MIPS: a database for genomes and protein sequences, *Nucleic Acids Res.* **28** (2000), 37–40.
30. Peterson, C.L. and Herskowitz, I.: Characterization of the Yeast SWI1, SEI1, and SWI3 Genes, which encode a global activator of transcription, *Cell* **68** (1992), 573–583.
31. Li, Y., Bjorklund, S., Jiang, Y.W., Kim, Y.J., Lane, W.S., Stillman, D.J. and Kornberg, R.D.: Yeast global transcriptional regulators Sin4 and Rgr1 are component of mediator complex/RNA polymerase II holoenzyme, *Proc. Natl. Acad. Sci. USA* **92** (1995), 10864–10868.
32. Tzamarias, D. and Struhl, K.: Functional dissection of the yeast Cyc8-Tup1 transcriptional co-repressor complex, *Nature* **369** (1994), 758–761.
33. Efron, B. and Gong, G.: A leisurely look at the bootstrap, the jackknife and crossvalidation, *Amer. Statistician* **37** (1982), 36–48.