JOURNAL OF BIOSCIENCE AND BIOENGINEERING
Vol. 96, No. 5, 421–428. 2003

Clustering Gene Expression Pattern and Extracting Relationship in Gene Network Based on Artificial Neural Networks

JIHUA HUANG,[1] HIROSHI SHIMIZU,[2] AND SUTEAKI SHIOYA[1]*

Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan[1] *and Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, Osaka University, Suita, Osaka 565-0871, Japan*[2]

Received 7 January 2003/Accepted 30 July 2003

Massive datasets such as gene expression profiles are accumulating along with the development of DNA microarray technologies. In this paper, we focus on mining biological relevant information such as typical expression patterns and the interconnections of gene networks from massive datasets. At first, the algorithm of a self-organizing map (SOM) was used to cluster gene expression data. Then, for the typical patterns extracted by the SOM, a three-layer artificial neural network (ANN) model was used to extract the relationships between the expression patterns. In order to evaluate the clustering analysis based on the SOM, biological and statistical indices were introduced. To validate the efficiency of the scheme proposed for extracting the relationships between the expression patterns with the ANN, a test dataset was created and used for the test. Finally, the interconnections of a typical pattern of early G1, late G1, S, G2, and M phases in a yeast cell cycle were extracted and visualized.

[**Key words:** clustering, gene expression pattern, self-organizing map, artificial neural network]

Gene expression monitoring on a genome-wide scale was first successfully demonstrated in yeast (1–4). Gene expression experiments can be repeated an arbitrary number of times to monitor the expression of different cell types, or the same cells under different conditions. Experiments have now been performed on many other organisms, including humans (5), mice (6), worms (7), and *Escherichia coli* (8). Genome-wide expression information is principally generated by three technologies: high-density oligonucleotide arrays (also called GeneChips) (9), cDNA microarrays (10), and serial analysis of gene expression (SAGE) (11). The three expression technologies all potentially give rise to a new genome-scale dataset and a further challenge in bioinformatics.

The first major bioinformatic task related to gene expression is analyzing those datasets derived from DNA array experiments. Expression data analysis can be divided into two parts. In the first part, the numerical structure of the expression data should be analyzed by a clustering method without biological knowledge such as the molecular mechanism. The investigation of the first part is performed by several methods such as hierarchical clustering (12, 13), self-organizing maps (SOMs) (14, 15), simulated annealing (16), and *k*-means (17).

The second part of expression data analysis is concerned with the extraction of the interconnections of mutual gene or gene clusters (18, 19). Several methods have been proposed for the second part (17, 20, 21). However, these two parts of studies were sometimes independently performed and discussed. In this paper, a two-step method was developed with an SOM model and an artificial neural network (ANN). The SOM was used for clustering and the ANN was used for extraction of the interconnections of gene clusters.

In order to evaluate the cluster analysis, biological and statistical indices were utilized. The efficiency of the scheme proposed for analyzing gene networks with the ANN was validated by a test dataset, in which the correct relationships of the gene expression patterns were known *a priori*. As a practical example, the previous study of Cho *et al.* (3) published at the website (http://genomics.stanford.edu) was adopted. The results indicated that our scheme was adequate.

MATERIALS AND METHODS

Analyzed data set and two neural networks used The yeast cell cycle data set provided by Cho *et al.* (3) contains time-course expression profiles for more than 6000 genes, with 17 time points for each gene taken at 10-min intervals covering nearly two yeast cell cycles (160 min). This data set was used as a practical example to analyze after excluding the data at the 90-min time point because of difficulties with scaling of the data. Therefore, 16 time points of data were actually employed.

A yeast genetic network between all gene expression patterns (more than 6000 genes) has been devised by using networks (20–22). Although it undoubtedly provided richer information in the genetic network covering all genes, it was generally time consuming, costly, and labor-intensive.

In this study, a two-step method with a neural network combination consisting of an SOM and a three-layer ANN was used to analyze the relationships of the yeast gene expression patterns in the

* Corresponding author. e-mail: shioya@bio.eng.osaka-u.ac.jp
phone/fax: +81-(0)6-6879-7444

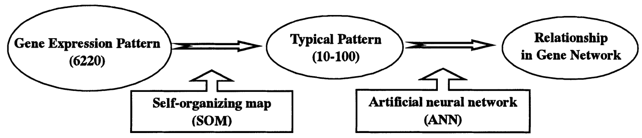
Gene Expression Pattern (6220)
Typical Pattern (10-100)
Relationship in Gene Network
Self-organizing map (SOM)
Artificial neural network (ANN)

FIG. 1. The proposed scheme with a neural network combination consisting of an SOM and a three layer ANN. At the first step, gene expression patterns were clustered to typical patterns by the SOM. Then, the interconnections in typical clusters were extracted by the ANN.

cell cycle as shown in Fig. 1. At the first step, thousands of gene expression patterns were clustered to a much smaller number of groups, namely, typical patterns were extracted by comparing their similarities of the gene expression data with the SOM. The interconnections of the typical clusters were then extracted by the three-layer ANN at the second step.

Data preprocessing Before clustering with the SOM, a variation filter was applied to eliminate the genes whose expression levels were relatively low and the genes that did not show significant changes during the time course. Namely, the following conditions were satisfied in the data used for analysis: (i) the absolute value of expression at all 16 time points was equal to or greater than 100 (in units used in the downloaded file) and (ii) at least greater than a threefold change in expression level during the time course was given. The profiles for the 1019 genes that passed the variation filter were normalized so that the expression level for each gene varied between zero and one.

Clustering with SOM The SOM is one of the best known unsupervised neural learning algorithms (23). The SOM layer is a low dimensional array of neurons (generally 2-D). An example of a 2-D SOM layer is shown in Fig. 2. In this paper, square SOMs (3×3, 5×5, ... 15×15) were employed. The dimension of the map is n by n neurons or n^2 of neurons in total ($i=1$ to n^2), which means n^2 of typical expression patterns will be extracted by clustering analysis with the SOM. The k-dimension of the vector was defined as an input vector $\boldsymbol{X}_j$ $[k\times 1]$, where j and k are the number of genes in the dataset ($j=1, 2, 3, \ldots, 1019$) and number of time points ($k=16$), respectively. Each element in the input layer, $\boldsymbol{X}_j$ is fully connected to n^2 neurons in the SOM layer. The neurons in the map of the SOM layer have vectors with the same dimension, 16 as that of the input vector $\boldsymbol{X}_j$, called weight vectors $\boldsymbol{W}_i$ ($i=1$ to n^2). In the figure, a winning neuron and its first neighbors are also shown.

At iteration of t in the training, a distance $\mathrm{d}(\boldsymbol{X}_j, \boldsymbol{W}_i)$ is defined and used for the measure of similarity between $\boldsymbol{X}_j$ and $\boldsymbol{W}_i$. In this paper, Euclidean distance metric was used. The neuron closest to the input vector is selected as a winning neuron with the corresponding weight vector, $\boldsymbol{W}_\mathrm{c}$, where c is the index of the winning neuron as Eq. 1. When the winning neuron is selected, the weights of the neurons at iteration of $(t+1)$ with two parameters of $N_\mathrm{c}(i, t)$ and α are updated by Eq. 2.

C(t) satisfies

$$\|\boldsymbol{X}_j(t)-\boldsymbol{W}_\mathrm{c}(t)\| = \min\{\|\boldsymbol{X}_j(t)-\boldsymbol{W}_i(t)\|\} \quad (1)$$

$$\boldsymbol{W}_i(t+1)=\boldsymbol{W}_i(t)+\alpha(t)N_\mathrm{c}(i,t)[\boldsymbol{X}_j(t)-\boldsymbol{W}_i(t)] \quad \text{for } i\le N_\mathrm{c}(i,t) \quad (2)$$

$$\boldsymbol{W}_i(t+1)=\boldsymbol{W}_i(t) \quad \text{for all other indices of } i \quad (3)$$

where $N_\mathrm{c}(i, t)$ is a parameter which indicates neurons as neighborhoods around the winning neuron. In this case, the neurons that belong to $N_\mathrm{c}(i, t)$ are updated, while the other neurons are not updated as in Eq. 3. The parameter $\alpha(t)$ is the learning rate, which is a monotonically decreasing function of t, and $N_\mathrm{c}(i, t)$ that denotes the neighborhood size of the winning neuron c at iteration of t. Usually the function α is decreased nonlinearly from close to unity towards zero. The neighborhood $N_\mathrm{c}(i, t)$ is also decreased as the number of iterations progress. In this paper, the SOM was implemented using the Matlab Neural Network toolbox (Mathworks, Natick, MA, USA).

Evaluation method of the number of clusters in the SOM The most important problem in application of the SOM to clustering gene expression profile pattern data is how to determine the number of clusters. In this study, the SOMs with 9 (3×3), 16 (4×4), 25 (5×5), 36 (6×6), 49 (7×7), 64 (8×8), 81 (9×9), 100 (10×10) and 225 (15×15) neurons were investigated to cluster the cell cycle dataset consisting of 1019 genes. In order to evaluate the performance of the clustering analysis, we referred to the previous results reported by Cho *et al.* (3), in which the fundamental patterns of gene expression with respect to yeast cell cycle were well understood from the biological view. Figure 3 shows the gene expression patterns of Early G1 (EG1), Late G1 (LG1), S, G2, and M phases of the cell cycle, which were reported by Cho *et al.* (3). The parenthetic number shows one of the gene members passed by our variation filter. These gene members were employed as "a test

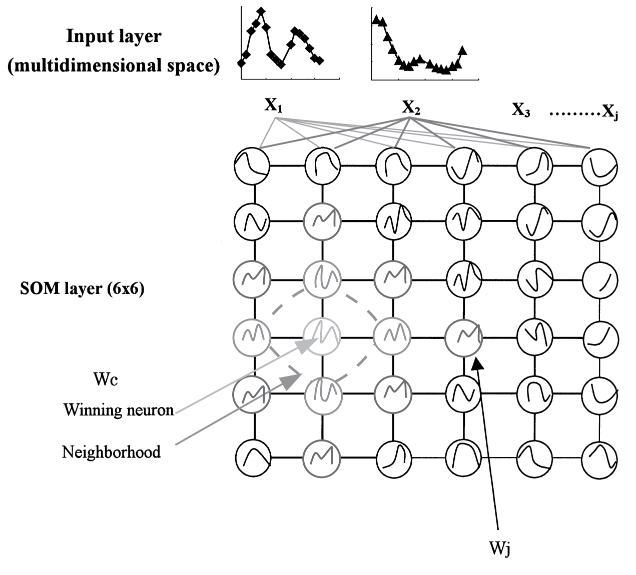

FIG. 2. The structure of the self-organizing map (SOM).

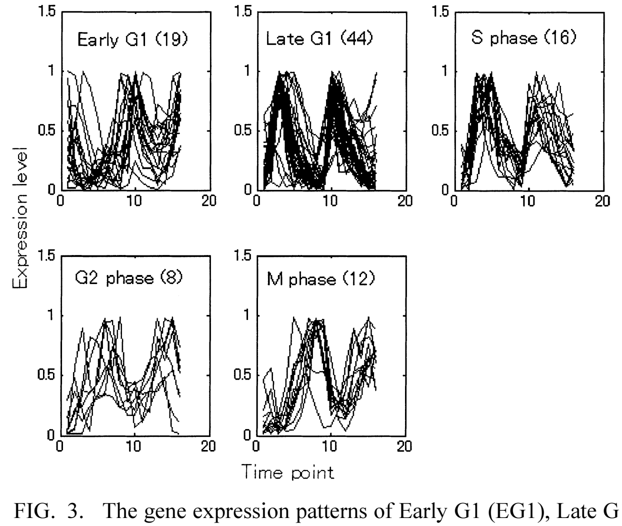

FIG. 3. The gene expression patterns of Early G1 (EG1), Late G1 (LG1), S, G2 and M phases of cell cycle, which were reported by Cho *et al.* (3). Beside the name of each cell cycle phase, the parenthetic number shows one of the gene members passed our variation filter.

dataset" or "a correct answer" of a gene in the each cell cycle phase to assess each SOM model.

In the evaluation of the SOM, a neuron was termed "a typical neuron" for each phase under consideration, in which most of the gene members of a particular phase of the cell cycle in the test dataset exist. The neurons specified as the "typical neuron" and its first neighbors were termed a "typical group". The number of clusters in the SOM was determined by evaluating biological and statistical indices shown as below.

Biological indices (CAP and FPP) Correct answer proportion (CAP) and false positive proportion (FPP) were defined to evaluate the clustering analysis. The CAP was a proportion of the number of gene members located in "the typical group" to the number of all the genes designated as "the correct answer" in each phase. FPP was defined as a proportion of the number of gene members located in all "other typical neurons" to the total genes designated as "the correct answer". In general, along with an increase in the number of neurons in the SOM, the values of both the CAP and FPP tend to become small. It is necessary to assess both the CAP and FPP when the SOM model is evaluated. In this study, the average of CAP and FPP in five cell cycle phases were employed to evaluate each SOM model. The values of the CAP and FPP were called biological indices in this study since they referred to the reported real expression pattern of the EG1 to M phase in the cell cycle as criteria.

Statistical indices (ASD, ARE, SmiW and AIC) used in evaluation Furthermore, each SOM model was evaluated by four statistical indices. The average of standard deviation (ASD) is the average of standard deviations of the data which belongs to the same neuron at each sampling time. The average of residual errors (ARE) is the average of distances between individual input data of and weight vector of a neuron after termination of the learning procedure.

The third index was defined as the similarity of the neurons with the neighborhood neurons. In general, to an excessive number of neurons in the SOM, it is expected the correlation between a neuron and its first neighbors will be large. Here SmiW was defined as a proportion of the number of pairs, for which the value of correlation coefficients between a neuron and its first neighbors exceeded 0.9, to the total pairs of a neuron with its first neighbors.

The Akaike information criterion (AIC) (24, 25) was also used to assess the SOM models. The AIC can be represented in Eq. 4 as

$$\mathrm{AIC} = \mathrm{N} \cdot \log \mathrm{ARE} + 2 \cdot \mathrm{P} \tag{4}$$

where N, ARE, and P stand for the number of sample data analyzed, the average of residual errors of the model (ARE, defined previously), and the number of parameters in the model, respectively. The first term in Eq. 4 reflects the accuracy of the model, and the second is its complexity or redundancy of parameters. In general, the best model should be selected, which has the smallest AIC value, indicating an optimum balance of accuracy and complexity.

Extracting interconnections of typical gene clusters by ANN After the typical expression patterns of the EG1, LG1, S, G2, and M phases were extracted by the SOM, the weight vectors of typical neurons were used as input data for an ANN model to analyze the interconnectivity and mutual regulation in terms of the characteristic shapes of expression profiles about the EG1, LG1, S, G2, and M phases.

ANNs are sophisticated techniques capable of modeling. The theory behind ANN has been sufficiently described in monographs and various papers (26–29). The ANN used in this work consists of three layers; input, hidden and output, as shown in Fig. 4, and sigmoid functions were used to transform the neuron output shown in Eqs. 5 and 6 as

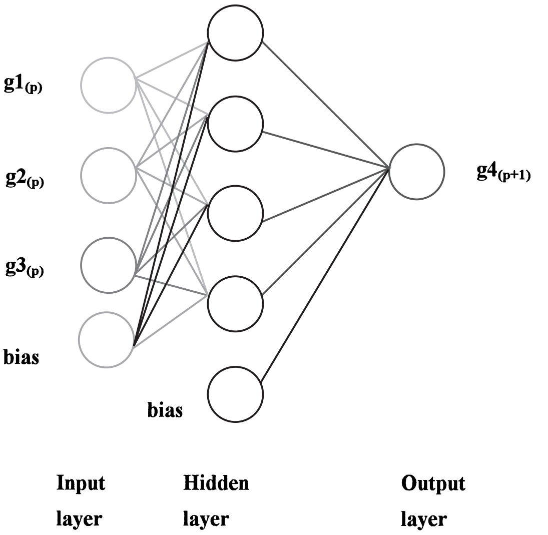

FIG. 4. The structure of the artificial neural network (ANN).

$$f(u) = \frac{1 - e^{-u}}{1 + e^{-u}} \tag{5}$$

$$g(u) = \frac{1}{1 + e^{-u}} \tag{6}$$

The number of neurons in the output layer was set at one. The numbers of neurons in the hidden layer was set to be only one greater than those in the input layer. The number of neurons in the input layer varied and depended on the number of input patterns analyzed. Weight vectors of the typical patterns of neurons in the SOM were used for dataset for the input and output layers in the training of the ANN. The back propagation method was used for the learning procedure (27).

One of the weight vectors of a typical neuron which was designated as "one of the specific phases for the cell cycle (one of the EG1, LG1, S, G2, and M phases)" by the SOM was used as output test data for the learning procedure of the ANN. This gene cluster was regarded as "a regulated gene candidate" by the different gene clusters. The weight vector of some other typical neuron which was designated as "different specific phases of the cell cycle" by the SOM was used as an input test data. This gene cluster was regarded as "gene candidate" which regulate the above gene.

First, the weight vector at a time of $(p+1)$ of the candidate regulated gene clusters was put into the neuron in the output layer of the ANN. Second, weight vectors at a current time point of p were put into the neurons in the input layer. Then, the interconnectivity between the gene expression patterns put into the neurons in the input and output layers was evaluated by assessing the training errors of the ANN. The performance index J was defined as the summation of squared errors (SSE) between the values of trained data Re (elements of weight vectors of the SOM) and the outputs Ot of the trained ANN as

$$J = \mathrm{SSE} = \sum_{l=1}^{n} (\mathrm{Ot}_l - \mathrm{Re}_l)^2 \tag{7}$$

where l and n are the time point of data and number of the total time point, respectively. It is expected that the gene expression pattern in the output layer can be well represented by the genes expression pattern in the input layer when there are interconnections between the two gene clusters. In such cases, the residual errors SSE is expected to be small. On the other hand, the SSE becomes large when there is no interconnection between the two gene clusters. Taking this into consideration, the SSE was used for the evaluation of extracting the regulation mechanism in gene networks.

RESULTS AND DISCUSSION

Determining the number of clusters in SOMs

Figure 5 shows the CAP and FPP values in the several SOMs with sizes from (3×3) to (15×15). The horizontal, right, and left vertical axes were the number of neurons in the SOMs, the CAP (Fig. 5, triangles) and FPP (Fig. 5, circles), respectively. It was apparent that the values of the CAP of the SOMs with a size of (7×7) to (15×15) were significantly small only near 50%. On the other hand, the values of the FPP of the SOMs with a size of (3×3) and (5×5) were significantly large even if their CAP values were also large. Compared to these SOMs, the (6×6) SOM was considered to be a better candidate to represent the characteristics of a gene expression profile.

Figure 6A shows the values of the ASD and ARE, and Fig. 6B shows the SmiW of SOMs with a size from (3×3) to (15×15), respectively. It is expected that the good candidate model has a good precision evaluated by the small values of the ASD and ARE. However, a further increase in the SOM size makes only excessive neurons even though the ASD and ARE are small because the values of the SmiW become large. This means that the correlation of neighbor neurons becomes very tight. In this study, the gradients of the ASD, ARE, and SmiW in each SOM were used to assess the candidate models as shown in Fig. 6C and 6D. By observing the values of gradient of the ASD, ARE, and SmiW, it was apparent that the increasing rates of the gradients of the ASD, ARE, and SmiW became less at the (6×6) SOM. Thus, the (6×6) SOM with 36 neurons was evaluated as a good candidate among the SOMs investigated. It was consistent with the result based on biological indices.

The right vertical axis of Fig. 6B shows the value of the AIC of the SOMs with sizes from (3×3) to (15×15). The (10×10) SOM would be selected as a good model since its AIC value was the lowest among all the SOMs. It should be noted also that in the AIC of the weights of the first term of the criterion of the accuracy in Eq. 4 were much larger than those of the second term of complexity. In this study, the number of sample gene expression data analyzed (N) was much larger than the number of parameters in the model (P).

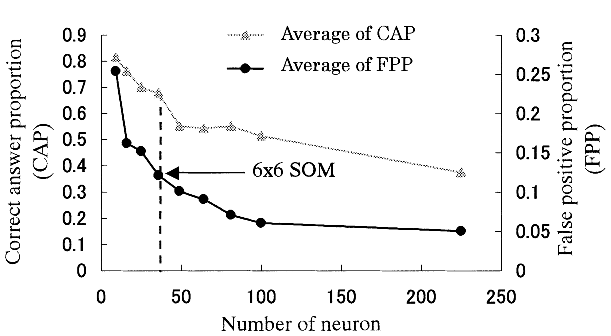

FIG. 5. The correct answer proportion (CAP) and false positive proportion (FPP) values in SOMs with sizes of (3×3) to (15×15). The horizontal, right, and left vertical axes are the number of neurons in SOM, CAP and FPP; triangles and circles represent CAP and FPP, respectively. The dashed line indicates the (6×6) SOM position.

The order of value of the first term in Eq. 4 was 10^4 and the second was in the range of 10^2 and 10^3. In such a case, it was not considered to be adequate that the criterion was adopted only based on the lowest AIC value. Here, the AIC gradient instead of the AIC value was used to assess the SOMs as shown in Fig. 6D, and then the (6×6) SOM was

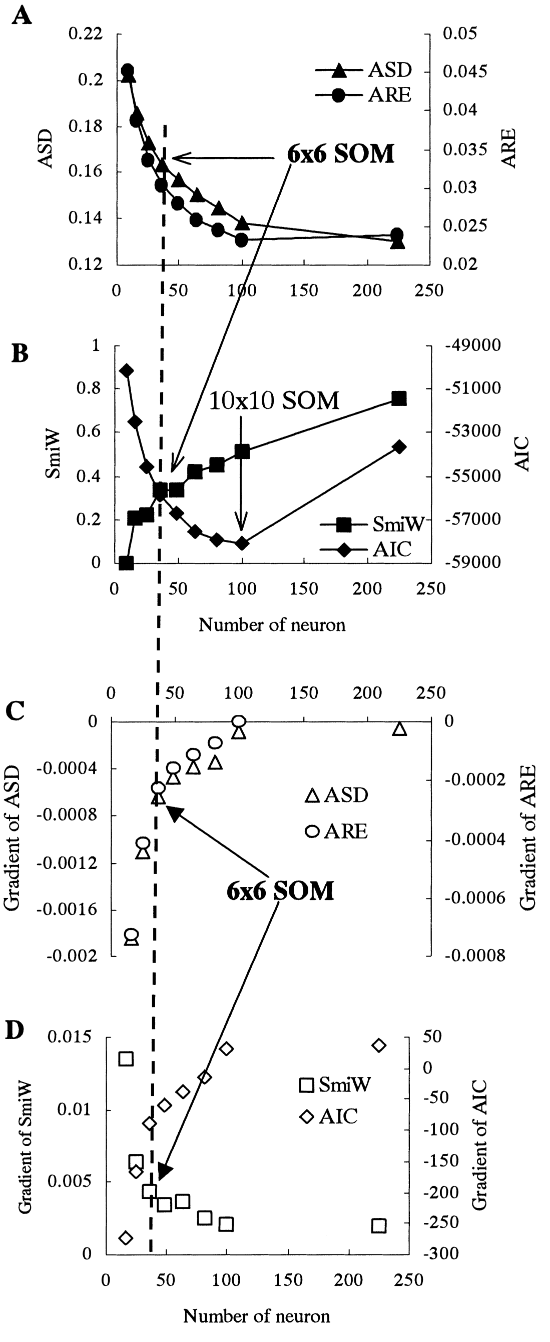

FIG. 6. The average of standard deviation (ASD) and average of residual errors (ARE) (A), and their gradients (C); average of similarity (SmiW) and Akaike information criterion (AIC) values (B), and their gradients (D) in SOMs. The horizontal axes are the number of neurons in SOMs. The left vertical axes are ASD (A), SmiW (B), gradient of ASD (C), and gradient of SmiW (D), while the right vertical axes are ARE (A), AIC (B), gradient of ARE (C), and gradient of AIC (D). The dashed line indicates the (6×6) SOM position.

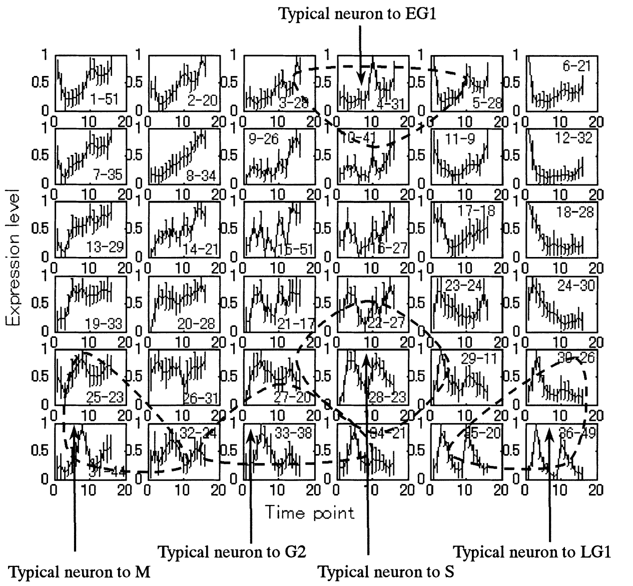

FIG. 7. The clustering result by the (6×6) SOM. Typical neurons to the EG1, LG1, S, G2, and M phase were clusters 4, 36, 28, 33, and 31 indicated by an arrowhead, respectively, while their typical groups are drawn by the dashed lines.

selected. This result was consistent with other results based on the CAP, FPP, ASD, ARE, and SmiW.

As an outcome, the (6×6) SOM was selected to represent the characteristics of the gene expression profile. Figure 7 shows the clustering result by the (6×6) SOM. The *i–m* written inside of the frame stands for the cluster number of *i* which involves *m* individual members of gene expression profiles. Those "typical neurons" to the EG1, LG1, S, G2, and M phases were cluster numbers of 4, 36, 28, 33 and 31 indicated by the arrowheads, respectively, while the "typical groups" are drawn by dashed line in Fig. 7. In this model, the CAP of EG1, LG1, S, G2, and M phase were 52.6%, 77.3%, 62.5%, 62.5% and 83.3%, while the FPP of each phase was 15.8%, 11.4%, 18.8%, 12.5% and 0%, respectively. The means of CAP and FPP were 67.7% and 11.7%, respectively.

Extracting the relationships in gene networks via ANN To validate the efficiency of the ANN method proposed for analyzing gene networks, a test dataset was artificially created, in which the correct interconnections of each gene expression patterns were artificially set as shown in Fig. 8. The g1, g2, g3, and g4 were step functions, the quadratic function of g1, combination of the sine curve of g2 and the second the linear function of g1, and the logarithmic function of g3, respectively. It was apparent that the g2, g3, and g4 changed according to g1, g1 and g2, and g3, respectively. The arrowheads in Fig. 9 indicate these relationships for each pattern.

The ANN models for all combinations of the genes from the g1 to g4 were tested by setting all the combinations of genes into inputs and outputs. All of the cause-and-effect relationships were evaluated by assessing the SSE as shown in Fig. 10. The vertical axis shows all the combinations of the relationships between the inputs and outputs, and the horizontal axis shows their SSE. Pairs of the g1 to g2, g2 to

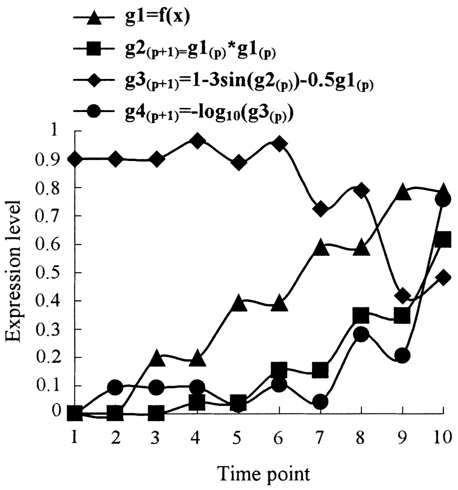

FIG. 8. A test dataset created, in which the correct connections of each gene expression patterns were artificially created.

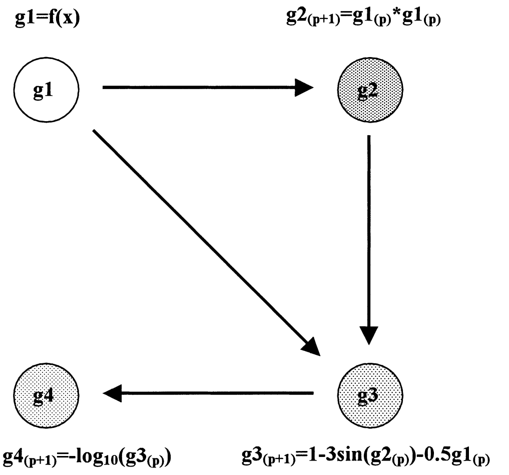
$g1=f(x)$
$g2_{(p+1)}=g1_{(p)}*g1_{(p)}$
g1
g2
g4
g3
$g4_{(p+1)}=-\log_{10}(g3_{(p)})$
$g3_{(p+1)}=1-3\sin(g2_{(p)})-0.5g1_{(p)}$

FIG. 9. Relationships between patterns of g1, g2, g3, and g4 in the test dataset created.

g1, g3 to g2 and g4 to g3 indicated by "#" represent the minimum values of the SSE among one-to-one variable combination genes. For example, in terms of the g4 as the output, the combination with the g3 as the input showed the minimum SSE among the all the pairs of one-to-one variable combinations, thus the g3 was recognized as the predominant conductor. Namely, the g3 was selected as the closest regulator to the g4. In this way, the closest regulator to each variable was selected. The value of 0.5-fold the mean of the SSE was employed as a threshold to evaluate the interconnections between variables, which is represented by the dashed line in Fig. 10. When the value of the SSE of the ANN for the one-to-one pair was smaller than the threshold value, the genes in the input layer were selected as regulators of the output layer. In test dataset, the g1, g1 and g2, and g3 were selected as regulators to g2, g3 and g4, respectively, which are indicated by asterisks in Fig. 10. These extracted relationships were completely the same as the correct answer artificially given (Fig. 9). Indirect connections among genes should be extracted, and the relationships be-

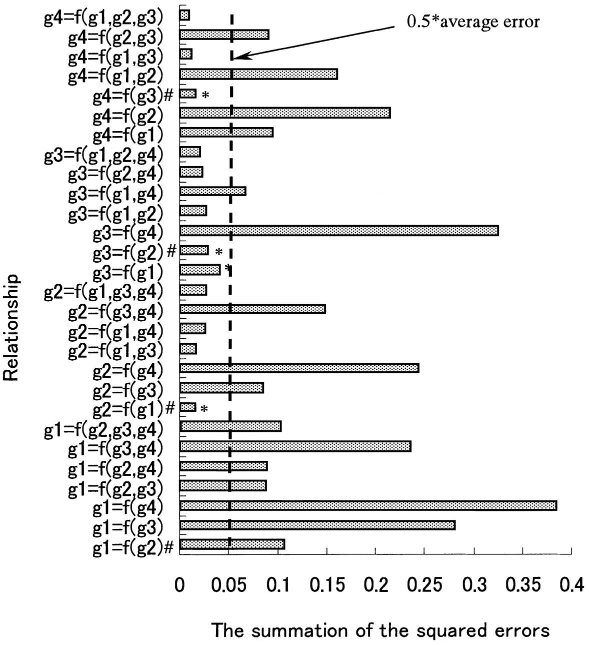

FIG. 10. The training SSE of all the cause-and-effect relationships combined by the g1, g2, g3, and g4. The vertical axis indicates relationships between a variable and other variable combination and the horizontal axis indicates training SSE. The symbol "#" indicates the minimum values among one-to-one variable combinations for a variable. Using the value of 0.5-fold the mean of the SSE as threshold, the g1, g1 and g2, and g3 were selected as regulators to g2, g3 and g4, respectively, which are indicated by asterisks.

tween more complicated combinations of one-to-two, three genes and threshold setting must be further studied in the future.

The proposed ANN model was applied to the extraction of interconnections among the typical expression patterns of the EG1, LG1, S, G2, and M phases represented by their "typical neurons". Figure 11 shows the SSE of all the expression pattern combinations and values of evaluating thresholds with 0.5-, 1.0-, and 1.5-fold the mean of the SSE. When the minimum value of the SSE was employed, the gene clusters of the EG1, LG1, S, G2, and M phases were selected as the dominant regulators to the EG1, LG1, S, G2, and M phases, respectively, and for those the symbol of "#" is attached in Fig. 11. When a value of 1.5-fold the mean of the SSE was used in the evaluation, combinations indicated by asterisks in Fig. 11 were selected as the interconnections of gene clusters with respect to the cell cycle. Except for the LG1 phase, other gene clusters were connected each other. The exception with respect to the LG1 was because that the typical pattern of the LG1 phase, which was represented by cluster 36 of the (6×6) SOM, does not significantly resemble other typical patterns as shown in Fig. 7.

In this method, since the interconnections of gene clusters were determined by a threshold, based on the SSE it depended on the value of the threshold, therefore, the number of the interconnections of the gene clusters in the network was completely dependent on the threshold. Figure 12B and 12C shows two patterns of interconnections of gene networks determined by the threshold value at 1.0- and 0.5-fold the mean of the SSE, respectively. It was apparent that

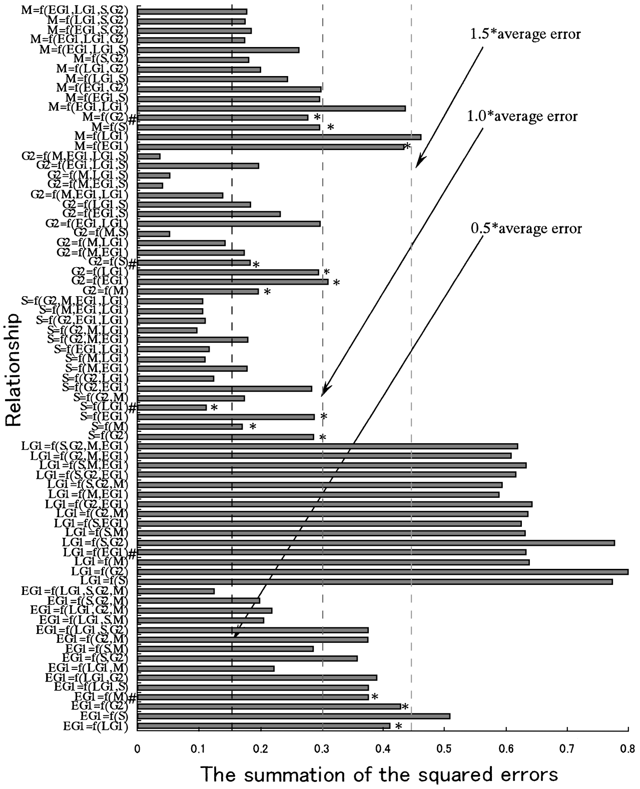

FIG. 11. The training SSE of all the cause-and-effect relationships combined by typical patterns of the EG1, LG1, S, G2, and M phases. "#" indicates the minimum values among one-to-one variable combinations for a variable. Using the value of 1.5-fold the mean of the SSE as a threshold, those pairs indicated by asterisks were selected as the closest partners with respect to regulation.

the number of the links of the network decreased along with a decrease in the complexity factor, and only one link remained when 0.5-fold the mean of the SSE was used as the threshold. More experiments using test datasets with various complexities of patterns should be run to determine an appropriate complexity factor. In addition, indirect connections among patterns are also necessary to discuss by means of assessing the SSE of one-to-two and one-to-three variables combinations. Here, for simplicity, the network shown in Fig. 12A was suggested to be an equivalent of representing a regulated relationship of gene expression with respect to the yeast cell cycle. To confirm this network extracted by the ANN, further investigation using wet biological experiments and other modeling algorithms should be considered. These investigations will be performed in the future.

In this paper, several sizes of SOMs were used to cluster gene expression data. To evaluate the clustering analysis, biological and statistical indices were used. As an outcome, the (6×6) SOM was selected as representing the characteristics of a gene expression profile with respect to a yeast cell cycle example. It has been reported that one important problem in the learning of the SOM is that the order of the learning of the data might significantly affect the learning results of the SOM. An improvement algorithm avoiding this problem was also proposed (30). In this paper, such an improvement was not undertaken. In this sense, the SOM algorithm

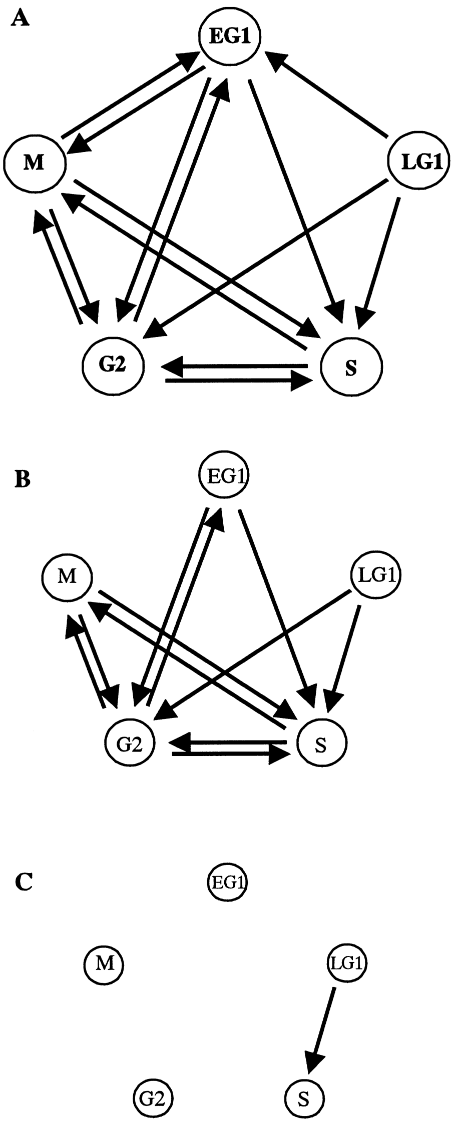

FIG. 12. Three networks of typical gene clusters related to cell cycle were determined by thresholds at the value of 1.5-fold (A), 1.0-fold (B) and 0.5-fold (C) the mean of the SSE, respectively.

should be further improved. However, the value of the CAP in the (6×6) SOM was close to 70% and genes recognized as the genes related to the particular cell cycle phase here are worth investigating.

For the typical patterns extracted by the SOM, the ANN was used to extract the relationships between the expression patterns. The efficiency of the algorithm was validated with a test dataset created artificially. Finally, the interconnections of the typical patterns of the early G1, late G1, S, G2, and M phases in the yeast cell cycle were extracted. It is believed the algorithm may be a helpful tool for clustering gene expression profile data and extracting their network, though further validation using other datasets is necessary as well. The two-step method with an SOM and an ANN was especially valuable for this method.

ACKNOWLEDGMENTS

This work was performed as a part of the Project for Development of a Technological Infrastructure for Industrial Bioprocesses of the New Energy and Industrial Technology Development Organization (NEDO) of the Ministry of Economy, Trade and Industry of Japan. This research was also supported in part by "The 21st Century Center of Excellence (COE) Program" of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

NOTATION

c : index for winning neurons in SOM
$f(u)$: nonlinear sigmoid function in ANN
$g(u)$: nonlinear sigmoid function in ANN
i : number of neurons in SOM
j : number of genes in input layer of SOM
J : evaluation index based on SSE
k : dimension of the input vector $\boldsymbol{X}_j$ in input layer of SOM
l : code of time point input into output layer of ANN
n : number of time point input into output layer of ANN
N_c : neighborhood size in SOM
Ot : output value of neuron in ANN
Re : real value of gene expression pattern
SSE : summation of squared errors on training ANN
t : training iteration in SOM
u : output of the neurons in ANN
$\boldsymbol{W}$: weight vectors of neurons in SOM
$\boldsymbol{X}$: input vector in input layer of SOM
$\alpha(t)$: learning rate in SOM

REFERENCES

1. **DeRisi, J. L., Iyer, V. R., and Brown, P. O.:** Exploring the metabolic and genetic control of gene expression on a genomic scale. Science, **278**, 680–686 (1997).
2. **Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I.:** The transcriptional program of sporulation in budding yeast. Science, **282**, 699–705 (1998).
3. **Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W.:** A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell, **2**, 65–73 (1998).
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 *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell, **9**, 3273–3297 (1998).
5. **Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O.:** The transcriptional program in the response of human fibroblasts to serum. Science, **283**, 83–87 (1999).
6. **Lee, C. K., Klopp, R. G., Weindruch, R., and Prolla, T. A.:** Gene expression profile of aging and its retardation by caloric restriction. Science, **285**, 1390–1393 (1999).
7. **Reinke, V., Smith, H. E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S. J., Davis, E. B., Scherer, S., Ward, S., and Kim, S. K.:** A global profile of germline gene expression in *C. elegans*. Mol. Cell, **6**, 605–616 (2000).
8. **Richmond, C. S., Glasner, J. D., Mau, R., Jin, H., and**

Blattner, F. R.: Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res., **27**, 3821–3835 (1999).

9. **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.:** Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol., **14**, 1675–1680 (1996).
10. **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**, 639–645 (1996).
11. **Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K. W.:** Characterization of the yeast transcriptome. Cell, **88**, 243–251 (1997).
12. **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**, 14863–14868 (1998).
13. **Alon, U., Barkai, N., Notterman, D. A., Gish, K., 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**, 6745–6750 (1999).
14. **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**, 2907–2912 (1999).
15. **Toronen, P., Kolehmainen, M., Wong, G., and Castren, E.:** Analysis of gene expression data using self-organizing maps. FEBS Lett., **451**, 142–146 (1999).
16. **Lukashin, A. V. and Fuchs, R.:** Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. Bioinformatics, **17**, 405–414 (2001).
17. **Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M.:** Systematic determination of genetic network architecture. Nat. Genet., **22**, 281–285 (1999).
18. **Bucher, P.:** Regulatory elements and expression profiles. Curr. Opin. Struct. Biol., **9**, 400–407 (1999).
19. **Gerstein, M. and Jansen, R.:** The current excitement in bioinformatics — analysis of whole-genome expression data: how does it relate to protein structure and function? Curr. Opin. Struct. Biol., **10**, 574–584 (2000).
20. **Friedman, N., Linial, M., Nachman, I., and Pe'er, D.:** Using Bayesian networks to analyze expression data. J. Comput. Biol., **7**, 601–620 (2000).
21. **Noguchi, H., Hanai, T., Honda, H., Harrison, L. C., and Kobayashi, T.:** Fuzzy neural network-based prediction of the motif for MHC class II binding peptides. J. Biosci. Bioeng., **92**, 227–231 (2001).
22. **Imoto, S., Goto, T., and Miyano, S.:** Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression, p. 175–186. *In* Proceedings of the 7th Pacific Symposium on Biocomputing. The Finnish Artificial Intelligence Society and Helsinki University of Technology, Helsinki (2002).
23. **Kohonen, T.:** Self-organizing maps. Springer-Verlag, Berlin (1995).
24. **Akaike, H.:** A new look at the statistical model identification. IEEE Trans. Automatic Control, **AC-19**, 716–723 (1974).
25. **Kimura, H.:** Control and modeling. Keisoku to Seigyo, **37**, 228–234 (1998). (in Japanese)
26. **Huang, J., Shimizu, H., and Shioya, S.:** Data preprocessing and output evaluation of an autoassociative neural network model for online fault detection in virginiamycin production. J. Biosci. Bioeng., **94**, 70–77 (2002).
27. **Karim, M. N. and Rivera, S. L.:** Artificial neural networks in bioprocess state estimation. Adv. Biochem. Eng. Biotechnol., **46**, 1–33 (1992).
28. **Uozumi, N., Yoshino, T., Shiotani, S., Suehara, K., Arai, F., Fukuda, T., and Kobayashi, T.:** Application of image analysis with neural network for plant somatic embryo culture. J. Ferment. Bioeng., **76**, 505–509 (1993).
29. **Yi-Hong, Z., Kosola, A., Linko, S., and Linko, P.:** Neural network applications for fermentation control, p. 477–480. *In* Proceedings of the International Conference on Engineering Applications of Artificial Neural Networks (EANN '95). The Finnish Artificial Intelligence Society and Helsinki University of Technology, Helsinki (1995).
30. **Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., Mori, H., and Ikemura, T.:** Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. Gene, **276**, 89–99 (2001).