BIOINFORMATICS

Vol. 18 no. 8 2002
Pages 1073–1083

Analysis of expression profile using fuzzy adaptive resonance theory

Shuta Tomida, Taizo Hanai, Hiroyuki Honda and Takeshi Kobayashi*

Department of Biotechnology, School of Engineering Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Received on September 3, 2001; revised on January 1, 2002; accepted on February 27, 2002

ABSTRACT

Motivation: It is well understood that the successful clustering of expression profiles give beneficial ideas to understand the functions of uncharacterized genes. In order to realize such a successful clustering, we investigate a clustering method based on adaptive resonance theory (ART) in this report.

Result: We apply Fuzzy ART as a clustering method for analyzing the time series expression data during sporulation of *Saccharomyces cerevisiae*. The clustering result by Fuzzy ART was compared with those by other clustering methods such as hierarchical clustering, k-means algorithm and self-organizing maps (SOMs). In terms of the mathematical validations, Fuzzy ART achieved the most reasonable clustering. We also verified the robustness of Fuzzy ART using noised data. Furthermore, we defined the correctness ratio of clustering, which is based on genes whose temporal expressions are characterized biologically. Using this definition, it was proved that the clustering ability of Fuzzy ART was superior to other clustering methods such as hierarchical clustering, k-means algorithm and SOMs. Finally, we validate the clustering results by Fuzzy ART in terms of biological functions and evidence.

Availability: The software is available at http//www.nubio.nagoya-u.ac.jp/proc/index.html.

Contact: taizo@brs.kyushu-u.ac.jp

INTRODUCTION

The recent advances of genome-scale sequencing and array technologies have made it possible to monitor simultaneously the expression pattern of thousands of genes (Cho *et al.*, 1998, 2001; DeRisi *et al.*, 1997; Khodursky *et al.*, 2000; Lashkari *et al.*, 1997; Spellman *et al.*, 1998). The following step is to discover the useful information from the expression data. Nowadays, one of the most exciting challenges is a cluster analysis for genome-wide expression data from DNA microarray hybridization, since the successful clustering result may give researchers beneficial ideas to understand the functions of uncharacterized genes. Therefore, various clustering methods, e.g. hierarchical clustering (Eisen *et al.*, 1998), k-means algorithm (Somogyi, 1999) and self-organized maps (SOMs; Tamayo *et al.*, 1999), have been examined and used to elucidate the fundamental and/or characteristic expression pattern. The classification of cell line, especially human cancers (Alizadeh *et al.*, 2000; Perou *et al.*, 1999; Ross *et al.*, 2000; Scherf *et al.*, 2000), as well as the analysis of temporally expressed genes of *Saccharomyces cerevisiae* (Chu *et al.*, 1998; Eisen *et al.*, 1998) was examined using the hierarchical clustering. However, various shortcomings of hierarchical clustering for the study of gene expression were discussed (Tamayo *et al.*, 1999), e.g. lack of robustness. Self-organized clustering is a useful and powerful method in the case of classifying huge sets of disorderly data into some significant groups. The k-means algorithm clusters a given set of input patterns into k groups, where k should be previously defined heuristically. However, in the case of analyzing the expression data, there is not a definite answer to decide the optimal number of k. SOMs were applied to analyze expression data (Tamayo *et al.*, 1999), and several genes that regulate transcriptional control were discussed. However, it seemed difficult to describe explicitly the distinctive feature of each cluster.

In the present paper, we applied fuzzy adaptive resonance theory (Fuzzy ART) to analyze experimental expression data. Fuzzy ART has been introduced by Carpenter *et al.* (1991b), as a member of ART-networks (Carpenter and Grossberg, 1987a,b; Carpenter *et al.*, 1991a). ART is a kind of self-organized clustering, which clusters a given set of input patterns into some groups. One of the characteristics of Fuzzy ART is the use of a similarity parameter, which is called the vigilance parameter ρ, and then the resulting number of clustered groups depends

*To whom correspondence should be addressed. Present address: Laboratory for Applied Biological Regulation Technology, School of Bioresource and Bioenvironmental Science, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.

only on the similarity between all input patterns. Another one is a weight vector W. A weight vector W_j of cluster j is adjusted through the learning procedure and it shows the representative pattern of the cluster j. Therefore, we can easily understand the clustering result by weight vectors. In the present paper, we describe the application of Fuzzy ART to the analysis of gene expression data. In order to compare the clustering results, we also applied hierarchical clustering, k-means algorithm and SOMs to analyze the same expression data.

METHODS

Data preprocessing

In this study, we used the expression data from the study of Chu *et al.* (1998). In the paper, *S. cerevisiae* was synchronized by transferring them to the sporulation medium (SPM) at $t = 0$ to maximize the synchrony of sporulation. RNA was harvested at time $t = 0, 0.5, 2, 5, 7, 9$ and 11.5 h after transfer to SPM. Polyadenylated RNA was prepared by purification with oligo (dT) cellulose column, and its expression level was measured using microarray analysis. Each gene's mRNA expression level just before transfer to SPM was used as the control. Therefore, the expression ratio at time t, R_t, of each gene is defined as follows.

$$R_t = \frac{mRNA\ measured\ at\ time\ t}{mRNA\ measured\ just\ before\ transfer\ to\ SPM}$$

About 6100 genes of expression profiles are included in the data, which is available at http://cmgm.stanford.edu/pbrown/sporulation. We followed the same criteria as Chu *et al.* (1998) to select the genes that showed 2.2-fold changes of mRNA levels during sporulation.

Fuzzy ART algorithm

Figure 1 shows the concept of a Fuzzy ART. In Carpenter *et al.* (1991b), they refer to a clustered group by Fuzzy ART as a 'category'. We call it a 'cluster' as a normal clustering way, in this paper. Fuzzy ART includes the following; an input vector I_{gene}, a weight vector W_j of cluster j, a choice parameter $\alpha > 0$, a learning rate parameter $0 \leqslant \beta \leqslant 1$, a vigilance parameter $0 \leqslant \rho \leqslant 1$, a choice function T_j and a match function M_j. An input vector I_{gene} corresponds to seven dimensional vectors of each gene. For example, in the case of *SGA1* gene in Table 1, an input vector I_{gene} of *SGA1* is defined as follows.

$$I_{SGA1} = |\,0.06,\ 0.19,\ 0.05,\ 0.36,\ 0.65,\ 0.65,\ 0.86\,|$$

Competitive learning of Fuzzy ART is illustrated step by step as follows.

At first, the winner cluster is chosen by the choice function. For each input I_{gene}, the choice function T_j of each cluster j is defined as follows to indicate the similarity between I_{gene} and W_j based on W_j.

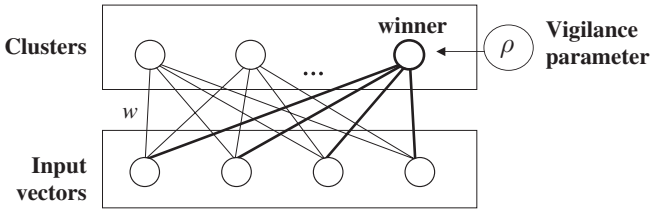

Fig. 1. Concept of Fuzzy ART.

$$T_j = \frac{|I_{\text{gene}} \wedge W_j|}{\alpha + |W_j|} \tag{1}$$

Where the minimum operator that is called 'and' operator in fuzzy theory, $\wedge$, is defined by

$$(x \wedge y)_{\text{i}} \equiv \min(x_{\text{i}}, y_{\text{i}}), \tag{2}$$

and where the operator $|\,x\,|$ is the sum of its components, $|\,x\,| = \sum x_{\text{i}}$. By a choice parameter α, the calculation error was prevented, if W_j becomes zero (Frank *et al.*, 1998). In the present study, α was set to 0.01. For example, when a weight vector W_{example} is defined as $W_{\text{example}} = |0.80,\ 0.75,\ 0.85,\ 0.70,\ 0.65,\ 0.50,\ 0.30|$, choice function T_{example} to I_{SGA1} is calculated as follows.

$$\begin{aligned}
|W_{\text{example}}| &= |0.80 + 0.75 + 0.85 + 0.70 + 0.65 + 0.50 + 0.30\,| \\
&= 4.55 \\
|I_{SGA1} \wedge W_{\text{example}}| &= |0.06 \wedge 0.80 + 0.19 \wedge 0.75 + 0.05 \wedge 0.85 \\
&\quad + 0.36 \wedge 0.70 + 0.65 \wedge 0.65 + 0.65 \wedge 0.50 + 0.86 \wedge 0.30| \\
&= |0.06 + 0.19 + 0.05 + 0.36 + 0.65 + 0.50 + 0.30\,| \\
&= 2.11 \\
T_{\text{example}} &= \frac{2.11}{0.01 + 4.55} \\
&= 0.46
\end{aligned}$$

The cluster j that has the maximal T_j is defined as the 'winner' cluster for input I_{gene}.

In the next step, the cluster selected above is judged to follow 'resonance' procedure or 'mismatch reset' procedure by the match function M_j defined by the following equation.

$$M_j = \frac{|I_{\text{gene}} \wedge W_j|}{|I_{\text{gene}}|} \tag{3}$$

'Resonance' procedure is carried out if the match function of the winner cluster for input I_{gene} is bigger than the vigilance criterionρ; that is expressed as

$$M_j \geqslant \rho \tag{4}$$

This means that the degree of similarity between the winner cluster and the current input I_{gene} is at least as high as vigilance. The match function indicates the similarity between I_{gene} and W_j based on I_{gene}. When 'resonance' procedure should be done, learning of the weight vector of winner cluster is performed. Learning of the weight vector W_j is updated according to the following equation.

$$W_j^{\text{new}} = \beta(I_{\text{gene}} \wedge W_j^{\text{old}}) + (1-\beta)W_j^{\text{old}} \quad (5)$$

Where β is a learning rate parameter. In this study, $\beta = 0.01$ was used in order to preserve the weight vector from being corrupted by noisy input patterns (Carpenter *et al.*, 1991b).

Otherwise, 'mismatch reset' procedure is carried out. That is expressed as

$$M_j < \rho \quad (6)$$

This means that the degree of similarity between the winner cluster and the current input I_{gene} is lower than the vigilance. When 'mismatch reset' procedure should be done, a new cluster that has the next maximal T_j is chosen by Equation (1) again.

When any cluster cannot satisfy Equation (4), a new cluster is generated according to the input vector I_{gene}. In the present study, we followed the Fast-commit slow-recode option (Carpenter *et al.*, 1991b). In this option, the learning rate parameter in Equation (5) is set to 1.0; $\beta = 1.0$. Initially, all of the elements of W_j^{initial} are set to 1.0; $(W_j^{\text{initial}})_{\text{i}} = 1.0$. Therefore, the weight vector of a new cluster is

$$\begin{aligned} W_j^{\text{newcluster}} &= \beta(I_{\text{gene}} \wedge W_j^{\text{initial}}) \\ &= I_{\text{gene}} \end{aligned} \quad (7)$$

These steps mentioned above are continued until every input vector I_{gene} belongs to any cluster.

In order to avoid a proliferation of categories, some methods have been proposed, see Carpenter *et al.* (1991b). In the present study, we used the complement coding to make stable clustering and to restrict the total number of the clusters. Briefly, the complement coding intends to recode the dimension of the input data. The complement of input I_{gene}, $I_{\text{gene}}^{\text{c}}$, are defined as follows.

$$(I_{\text{gene}}^{\text{c}})_{\text{i}} = 1 - (I_{\text{gene}})_{\text{i}} \quad (8)$$

In the case that the original input data are m-dimensional data, $I_{\text{gene}}^{\text{org}} = (I_1, I_2, \ldots, I_m)$, complement coded input I_{gene} is the 2m-dimensional data as follows.

$$\begin{aligned} I_{\text{gene}} &= (I_{\text{gene}}, I_{\text{gene}}^{c}) \\ &= (I_1, I_2, \ldots, I_m, 1-I_1, 1-I_2, \ldots, 1-I_m) \end{aligned} \quad (9)$$

Therefore, inputs preprocessed into the complement coding are automatically normalized.

Table 1. List of 45 genes

Name	ORF	Time of expression
CDC14	YFR028C	
CDC16	YKL022C	
CDC20	YGL116W	
CDC23	YHR166C	
CDC5	YMR001C	
DIT1	YDR403W	Mid–Late
DIT2	YDR402C	Mid–Late
DMC1	YER179w	Early
HOP1	YIL072W	Early
IME2	YJL106W	Early
IME4	YGL192W	Early
ISC10	YER180c	
MEI4	YER044c-a	Early
MEI5	YPL121C	
MEK1	YOR351C	Early
MPS1	YDL028C	
MSH4	YFL003C	
MSH5	YDL154W	
MSI1	YBR195C	
NDT80	YHR124W	
POL30	YBR088C	
RAD51	YER095w	
RAD54	YGL163C	
RAP1	YNL216W	
REC102	YLR329W	Early
REC104	YHR157W	Early
REC114	YMR133W	Early
RED1	YLR263W	Early
RFA1	YAR007C	
SAE3	YHR079C-B	
SGA1	YIL099W	Late
SPO11	YHL022C	Early
SPO12	YHR152W	Middle
SPO13	YHR014W	Early
SPO16	YHR153C	Early
SPO20	YMR017W	
SPR1	YOR190W	Late
SPR3	YGR059W	Late
SPR6	YER115c	
SPS1	YDR523C	Middle
SPS18	YNL204C	
SPS19	YNL202W	
YPT1	YFL038C	
ZIP1	YDR285W	Early
ZIP2	YGL249W	

The blank in 'time of expression' column means that no description for the induction period is mentioned in the Mitchell paper (1994).

RESULTS AND DISCUSSION

Result of data preprocessing

We extracted 522 genes as induced genes from the database. Among these 522 genes, we selected 45 genes that have some roles in meiosis and sporulation by Kupiec *et al.* (1997). Table 1 shows the 45 genes used for the following analysis. The column of 'time of expression'

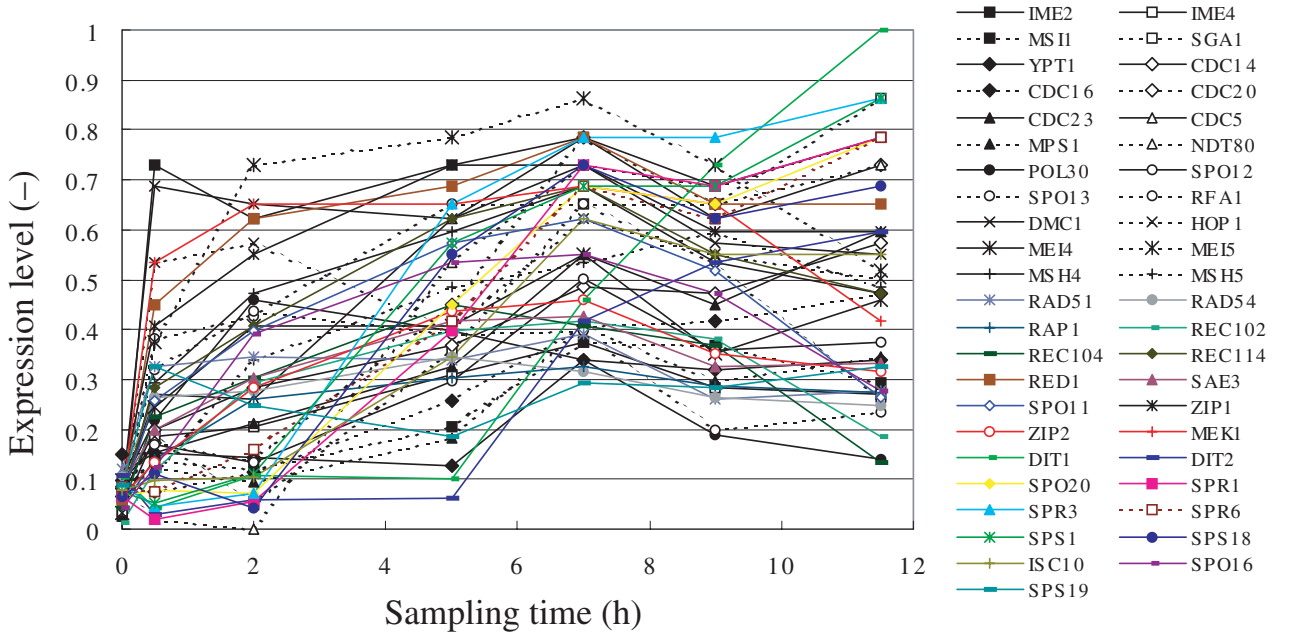

Fig. 2. Expression profiles of 45 genes used for this analysis.

shows the temporal phase, in which expression of each gene is induced, defined by Mitchell (1994). The blank in 'time of expression' column of Table 1 means that no description for the induction period is mentioned in Mitchell (1994).

Since expression ratios of each gene were measured at time $t = 0, 0.5, 2, 5, 7, 9$ and 11.5 h after transfer to SPM, each gene has seven dimensional data. The ranges of $\log_2 R_t$ of the selected 45 genes were from -0.78 to 6.65, and they were normalized from 0.0 to 1.0 to be used as the input data for the fuzzy operator in the following analysis. Figure 2 shows the expression profiles of 45 genes used for this analysis.

Identification of the optimal number of clusters for Fuzzy ART

At first, we investigated the effect of vigilance parameter on Fuzzy ART. The number of generated clusters increased when relatively higher vigilance parameter was applied. That is explained as follows. If the vigilance is relatively high, the 'mismatch reset' procedure is done easily and a new cluster is also generated more often. Figure 3 shows the number of clusters generated under various vigilance parameters. When the vigilance parameter was less than 0.86, the number of generated clusters was not so much affected by the vigilance parameter. With the range over 0.86, the number of clusters increased sharply. We examined four cases of the number of clusters such as 4, 5, 6 and 7, which corresponded to 0.83, 0.86, 0.867 and 0.87 of the vigilance parameter, respectively (Table 2).

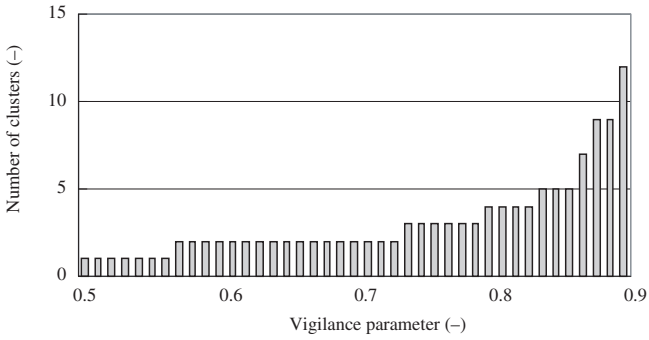

Fig. 3. Effect of vigilance parameter on the number of clusters generated by Fuzzy ART. Condition is a choice parameter $\alpha = 0.01$, a learning rate parameter $\beta = 0.01$.

In order to identify the optimal number of clusters, we defined the correctness ratio for the clustering result based on the previously reported biological results. The calculation for the correctness ratio was executed as follows. 'Early', 'Middle', 'Mid–Late' and 'Late' genes, which were characterized in Mitchell (1994), were used as 'index genes'. The majority of the index gene defined the character of the cluster. For example, in the case of the number of clusters 5, the 3rd cluster was defined not as 'Middle' but as 'Early' because four 'Early' genes were included in the 3rd cluster as the majority while one 'Middle' gene was included as the minority. The

Table 2. Comparison of clustering results using Fuzzy ART for 4, 5, 6 and 7 clusters

4 clusters Cluster	Name	Time	5 clusters Cluster	Name	Time	6 clusters Cluster	Name	Time	7 clusters Cluster	Name	Time
	DMC1	Early		*DMC1*	Early		DMC1	Early		*DMC1*	Early
1	*IME2*	Early	**1**	*IME2*	Early	**1**	IME2	Early	**1**	*IME2*	Early
	MEI5			*MEI5*			MEI5			*RED1*	Early
	RED1	Early		*RED1*	Early		RED1	Early	**2**	*MEI5*	
	HOP1	Early		*HOP1*	Early		HOP1	Early		*ZIP1*	Early
	MEK1	Early		*MEK1*	Early		MEK1	Early		*HOP1*	Early
	MSH4			*MSH4*			MSH4			*MEK1*	Early
	MSH5			*MSH5*		**2**	REC114	Early	**3**	*MSH4*	
2	*REC114*	Early	**2**	*REC114*	Early		SPO11	Early		*REC114*	Early
	SPO11	Early		*SPO11*	Early		SPO13	Early		*SPO11*	Early
	SPO13	Early		*SPO13*	Early		ZIP1	Early		*SPO13*	Early
	SPO16	Early		*SPO16*	Early		CDC14			*CDC14*	
	ZIP1	Early		*ZIP1*	Early		CDC23			*CDC23*	
	CDC14			*CDC14*		**3**	MEI4	Early	**4**	*ISC10*	
	CDC16			*CDC23*			MSH5			*MEI4*	Early
	CDC23			*IME4*	Early		SPO16	Early		*MSH5*	
	IME4	Early		*MEI4*	Early		IME4	Early		*SPO16*	Early
	MEI4	Early		*MPS1*			MPS1			*CDC16*	
	MPS1			*MSI1*			MSI1			*IME4*	Early
	MSI1			*POL30*			POL30			*MPS1*	
	POL30			*RAD51*			RAD51			*MSI1*	
	RAD51		**3**	*RAD54*			RAD54			*POL30*	
3	*RAD54*			*RAP1*			RAP1			*RAD51*	
	RAP1			*REC102*	Early	**4**	REC102	Early		*RAD54*	
	REC102	Early		*REC104*	Early		REC104	Early	**5**	*RAP1*	
	REC104	Early		*RFA1*			RFA1			*REC102*	Early
	RFA1			*SAE3*			SAE3			*REC104*	Early
	SAE3			*SPO12*	Middle		SPO12	Middle		*RFA1*	
	SPO12	Middle		*SPS19*			SPS19			*SAE3*	
	SPS19			*YPT1*			YPT1			*SPO12*	Middle
	YPT1			*ZIP2*			ZIP2			*SPS19*	
	ZIP2			*CDC16*			CDC16			*YPT1*	
	CDC20		**4**	*DIT1*	Mid–Late	**5**	DIT1	Mid–Late		*ZIP2*	
	CDC5			*DIT2*	Mid–Late		DIT2	Mid–Late	**6**	*DIT1*	Mid–Late
	DIT1	Mid–Late		*CDC20*			CDC20			*DIT2*	Mid–Late
	DIT2	Mid–Late		*CDC5*			CDC5			*CDC20*	
	ISC10			*ISC10*			ISC10			*CDC5*	
	NDT80			*NDT80*			NDT80			*NDT80*	
4	*SGA1*	Late		*SGA1*	Late		SGA1	Late		*SGA1*	Late
	SPO20		**5**	*SPO20*		**6**	SPO20		**7**	*SPO20*	
	SPR1	Late		*SPR1*	Late		SPR1	Late		*SPR1*	Late
	SPR3	Late		*SPR3*	Late		SPR3	Late		*SPR3*	Late
	SPR6			*SPR6*			SPR6			*SPR6*	
	SPS1	Middle		*SPS1*	Middle		SPS1	Middle		*SPS1*	Middle
	SPS18			*SPS18*			SPS18			*SPS18*	

The blank in "time" column means that no description for the induction period is mentioned in the Mitchell paper (1994).

majority genes included in each cluster, e.g. *IME4*, *MEI4*, *REC102* and *REC104* in the 3rd cluster, were defined as the correctly clustered genes while the minority, *SPO12*, was defined as the incorrectly clustered one. According to this definition, the correctness ratio was calculated as the ratio of correctly clustered genes among the 21 'index genes' shown in Table 1.

In the case of the number of clusters 4, 14 'Early' genes and three 'Late' genes were correctly clustered. It resulted that 17 genes among 21 'index genes' were correctly clustered. Therefore, the correctness ratio was 0.81 (Table 3). In the case of the number of clusters 5, 14 'Early' genes, two 'Mid–Late' genes and three 'Late' genes were correctly clustered and the assignment of two

Table 3. Comparison of correctness ratio using four clustering methods for 4, 5, 6 and 7 clusters

	Clutsers 4	Clusters 5	Clusters 6	Clusters 7
Fuzzy ART	0.81	0.90	0.90	0.90
Hierarchical clustering	0.81	0.81	0.90	0.90
k-means clustering	0.76	0.86	0.86	0.86
SOMs	0.86	0.86	0.86	0.86

'Middle' genes were incorrect. It resulted that 19 genes were correctly clustered, and the correctness ratio was 0.90. It is clear that the correctness ratio improved from 0.81 to 0.90 when the number of clusters was changed from 4 to 5, while it remained 0.90 when the number of clusters was changed from 5 to 7.

As shown in Table 2, the 4th cluster in the number of clusters 4 contains 'Middle', 'Mid–Late' and 'Late' genes, and this is not good biologically. In the case of the number of clusters 6, four clusters were generated for 'Early' genes, which seems not so good for a proper clustering. The 2nd and 6th clusters in the number of clusters 7 contain only two genes, which means so-called overclustering. Generating too many clusters makes it difficult to comprehend the simple features of the expression patterns. Therefore, in the present study, total 5 clusters seemed to be reasonable for the Fuzzy ART clustering.

Comparison of clustering results with those using other clustering methods

In order to compare the optimal number of clusters for other clustering methods, the correctness ratios were compared with other clustering methods such as hierarchical clustering, *k*-means algorithm and SOMs. The basic idea of hierarchical clustering is to assemble a set of 45 genes of seven dimensional expression data into a tree, where genes are joined by very short branches if they are very similar to each other, and by increasingly longer branches as their similarities decrease (Hartigan, 1975; Eisen *et al.*, 1998). We used the Peason correlation coefficient to define the similarity and the average linkage to assemble the items. In the analysis using hierarchical clustering, the number of the cluster can be chosen from one to the number of the data set. In the present study, we selected the same number as that of generated clusters using Fuzzy ART in order to compare the clustering results.

The *k*-means algorithm clusters a given set of input patterns into *k* groups (Frank *et al.*, 1998) and *k* was also set to the same number as that of Fuzzy ART. One-dimensional SOMs (Eisen *et al.*, 1998) were also used for the analysis, and 45-gene expression data were also classified into the same number as that of the generated clusters using Fuzzy ART. These three clustering methods are downloaded from http://rana.stanford.edu/clustering (Eisen *et al.*, 1998).

The correctness ratios using hierarchical clustering, *k*-means algorithm and SOMs are also shown in Table 3. It is clear that the correctness ratio of *k*-means clustering improved from 0.76 to 0.86 when the number of clusters was changed from 4 to 5, while it remained 0.86 when the number of clusters was changed from 5 to 7. In the case of hierarchical clustering, the correctness ratio improved from 0.81 to 0.90 when the number of clusters was changed from 5 to 6. In the case of SOMs, the correctness ratio remained 0.86 when the number of clusters was changed from 4 to 7. Therefore, three clustering methods, such as Fuzzy ART, *k*-means clustering and SOMs, achieved the highest correctness ratio at the number of clusters 5, although hierarchical clustering achieved it at the number of clusters 6. As mentioned above, generating too many clusters makes it difficult to comprehend the simple features of the expression patterns. Therefore, in the present study, we temporarily set the number of clusters to five.

Table 4 shows the clustering results using hierarchical clustering. As shown in Table 4, the 1st and 2nd clusters contain only one and two genes, respectively, and they were unchanged even if the number of clusters was changed from 4 to 7.

Table 5 shows the clustering results at the number of clusters 5 using *k*-means clustering and SOMs. As shown in Tables 2 and 5, each cluster contains several genes in the cases of Fuzzy ART, *k*-means clustering and SOMs. As shown in the above section, there was a cluster containing 'Middle', 'Mid–late' and 'Late' genes in the case of small numbers of clusters, while overclustering was observed in the case of a large number of clusters. Therefore, the cluster number 5 seemed to be sufficient and appropriate for clustering in the case of Fuzzy ART. A similar tendency was observed in the cases of *k*-means clustering and SOMs although the clustering results for the cluster number 5 are only shown in Table 5.

Therefore, in the present paper, we considered that Fuzzy ART is more superior to other clustering methods and we set the number of clusters to 5 in the following analysis.

Clustering results of Fuzzy ART

Figure 4 shows the weight vectors for each cluster, which are generated and updated through competitive learning using the following parameter: choice parameter=0.1, vigilance parameter=0.86, learning rate parameter=0.01. These vectors represent the profiles of genes included in each cluster. The cluster 1 includes the genes that are induced in the 'Early' phase of sporulation and high-expression level continues throughout sporulation. The expression levels of genes in the cluster 2 gradually in-

Table 4. Comparison of clustering results of hierarchical clustering using variable vigilance for 4, 5 and 7 clusters

4 clusters			5 clusters			6 clusters			7 clusters		
Cluster	Name	Time	Cluster	Name	Time	Cluster	Name	Time	Cluster	Name	Time
1	*SPS19*		**1**	*SPS19*		**1**	*SPS19*		**1**	*SPS19*	
2	*RFA1*		**2**	*RFA1*		**2**	*RFA1*		**2**	*RFA1*	
	POL30			*POL30*			*POL30*			*POL30*	
	DMC1	Early		*DMC1*	Early		*DMC1*	Early		*DMC1*	Early
	HOP1	Early		*HOP1*	Early		*HOP1*	Early		*HOP1*	Early
	IME2	Early	**3**	*IME2*	Early	**3**	*IME2*	Early	**3**	*IME2*	Early
	MEK1	Early		*MEK1*	Early		*MEK1*	Early		*MEK1*	Early
	RAD51			*RAD51*			*RAD51*			*RAD51*	
	RAD54			*RAD54*			*RAD54*			*RAD54*	
	IME4	Early		*IME4*	Early		*IME4*	Early		*IME4*	Early
	MEI4	Early		*MEI4*	Early		*MEI4*	Early		*MEI4*	Early
	MEI5			*MEI5*			*MEI5*			*MEI5*	
	MSH4			*MSH4*			*MSH4*			*MSH4*	
3	*RAP1*			*RAP1*			*RAP1*			*RAP1*	
	REC102	Early		*REC102*	Early		*REC102*	Early		*REC102*	Early
	REC104	Early		*REC104*	Early		*REC104*	Early		*REC104*	Early
	REC114	Early	**4**	*REC114*	Early	**4**	*REC114*	Early	**4**	*REC114*	Early
	RED1	Early		*RED1*	Early		*RED1*	Early		*RED1*	Early
	SAE3			*SAE3*			*SAE3*			*SAE3*	
	SPO11	Early		*SPO11*	Early		*SPO11*	Early		*SPO11*	Early
	SPO13	Early		*SPO13*	Early		*SPO13*	Early		*SPO13*	Early
	SPO16	Early		*SPO16*	Early		*SPO16*	Early		*SPO16*	Early
	ZIP1	Early		*ZIP1*	Early		*ZIP1*	Early		*ZIP1*	Early
	ZIP2			*ZIP2*			*ZIP2*			*ZIP2*	
	CDC14			*CDC14*			*DIT1*	Mid–Late		*DIT1*	Mid–Late
	CDC16			*CDC16*		**5**	*DIT2*	Mid–Late	**5**	*DIT2*	Mid–Late
	CDC20			*CDC20*			*YPT1*			*YPT1*	
	CDC23			*CDC23*			*CDC14*			*CDC14*	
	CDC5			*CDC5*			*CDC16*			*CDC16*	
	DIT1	Mid–Late		*DIT1*	Mid–Late		*CDC20*			*CDC20*	
	DIT2	Mid–Late		*DIT2*	Mid–Late		*CDC23*			*CDC23*	
	ISC10			*ISC10*			*CDC5*			*CDC5*	
	MPS1			*MPS1*			*ISC10*			*ISC10*	
	MSH5			*MSH5*			*MPS1*			*MPS1*	
4	*MSI1*		**5**	*MSI1*			*MSH5*			*MSI1*	
	NDT80			*NDT80*		**6**	*MSI1*		**6**	*NDT80*	
	SGA1	Late		*SGA1*	Late		*NDT80*			*SGA1*	Late
	SPO12	Middle		*SPO12*	Middle		*SGA1*	Late		*SPO12*	Middle
	SPO20			*SPO20*			*SPO12*	Middle		*SPO20*	
	SPR1	Late		*SPR1*	Late		*SPO20*			*SPR1*	Late
	SPR3	Late		*SPR3*	Late		*SPR1*	Late		*SPR3*	Late
	SPR6			*SPR6*			*SPR3*	Late		*SPR6*	
	SPS1	Middle		*SPS1*	Middle		*SPR6*			*SPS1*	Middle
	SPS18			*SPS18*			*SPS1*	Middle		*SPS18*	
	YPT1			*YPT1*			*SPS18*		**7**	*MSH5*	

The blank in 'time' column means that no description for the induction period is mentioned in the Mitchell paper (1994).

crease untill about 7 h and then decrease. The genes in the cluster 3 are not induced distinctly independent of temporal phase. The cluster 4 includes the genes with expression peak in 'Late' phase of sporulation. The genes included in the cluster 5 express strongly in latter phases and its expression level increases according to time passed.

Comparison of gap index

When we analyze a set of time series expression data, it seems no less necessary than important to consider the shapes of expression profiles not only as simply several dimensional inputs but also as the timely continuous data during a specific biological phase. In this point of view,

Table 5. Clustering results using k-means algorithm and SOMs for 5 clusters

k-means algorithm			SOMS		
Cluster	Name	Time	Cluster	Name	Time
1	*DMC1*	Early	1	*DMC1*	Early
	HOP1	Early		*HOP1*	Early
	IME2	Early		*IME2*	Early
	MSI1			*MEK1*	Early
	RAD51			*POL30*	
	RAD54			*RAD51*	
	RFA1			*RAD54*	
	SPS19			*REC104*	Early
2	*IME4*	Early		*RFA1*	
	MEI4	Early		*SPO13*	Early
	MEK1	Early		*SPS19*	
	MSH4		2	*IME4*	Early
	POL30			*MEI4*	Early
	RAP1			*MEI5*	
	REC102	Early		*MSH4*	
	REC104	Early		*RAP1*	
	REC114	Early		*REC102*	Early
	RED1	Early		*REC114*	Early
	SAE3			*RED1*	Early
	SPO11	Early		*SAE3*	
	SPO13	Early		*SPO11*	Early
	SPO16	Early		*SPO16*	Early
	ZIP1	Early		*ZIP1*	Early
	ZIP2			*ZIP2*	
3	*CDC14*		3	*CDC14*	
	CDC23			*CDC5*	
	MEI5			*MSH5*	
	MPS1			*MSI1*	
	YPT1			*SPO12*	Middle
4	*CDC20*		4	*CDC16*	
	CDC5			*CDC20*	
	MSH5			*CDC23*	
	SPO12	Middle		*ISC10*	
5	*CDC16*			*MPS1*	
	DIT1	Mid–Late		*SPS18*	
	DIT2	Mid–Late		*YPT1*	
	ISC10		5	*DIT1*	Mid–Late
	NDT80			*DIT2*	Mid–Late
	SGA1	Late		*NDT80*	
	SPO20			*SGA1*	Late
	SPR1	Late		*SPO20*	
	SPR3	Late		*SPR1*	Late
	SPR6			*SPR3*	Late
	SPS1	Middle		*SPR6*	
	SPS18			*SPS1*	Middle

we propose to analyze the similarity of profiles in terms of two-dimensional area, here axes of the two-dimensions are 'time' and 'expression level'. Then, we define the 'gap index' so as to evaluate the similarity of profiles as an area between each profile and average profile the during the temporal phase. An average profile for cluster n is defined as an average of all profiles of cluster n. The concept of gap is shown as a shaded area in Figure 5.

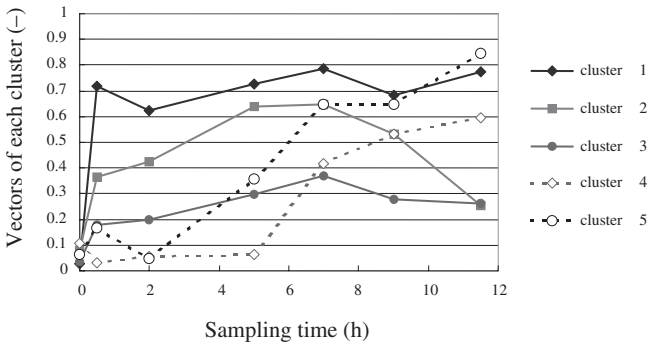

Fig. 4. Weight vector W of five clusters generated by Fuzzy ART.

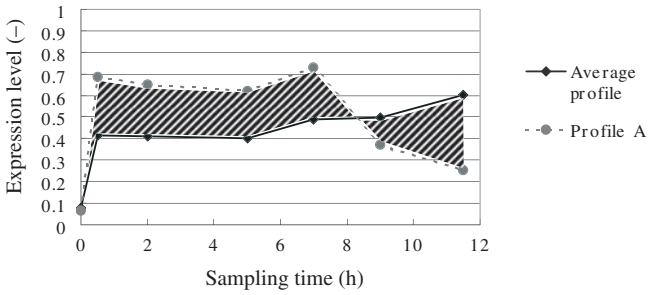

Fig. 5. Concept of gap index.

Figure 6 shows the gap index for each cluster in the four clustering methods mentioned above. For each cluster, the gap index of Fuzzy ART was set to 100 and the other gap indexes were calculated as a relative value against it. Since the cluster 1 in hierarchical clustering contains only one gene, the gap index of the cluster 1 in hierarchical clustering is void. It is clear that the average gap index of Fuzzy ART is the lowest.

Comparison of distribution of input profile

For further discussion, we compared the distribution of profiles clustered in the same cluster based on the standard deviation (SD). We define the average SD of each sampling time point t (ASD_t) as follows. At first, SD of each sampling time point t of cluster n ($\mathrm{SD}_{t,n}$) is calculated. Then, ASD_t is defined as an average of five $\mathrm{SD}_{t,n}$. Here n means the number of cluster ($n = 1, 2, 3, 4, 5$). ASD_t for each sampling time point t in four clustering methods were described in Figure 7. In addition, SD of 45 genes for each sampling time point t are also shown in Figure 7. This figure shows that ASD_2, ASD_5, ASD_7 and ASD_9 in Fuzzy ART are remarkably smaller than those in the other clustering methods. Especially, ASD_5 and ASD_7 of k-means algorithm and SOMs are almost the same to that of SD of 45 genes. This proves

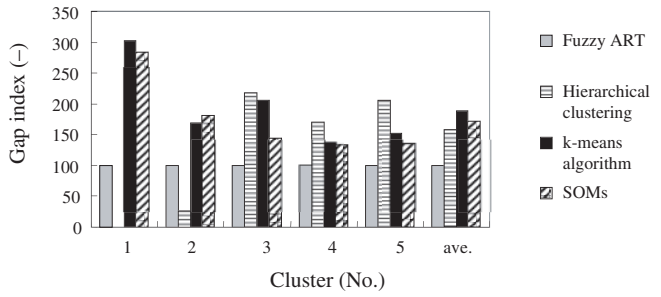

Fig. 6. Comparison of gap index.

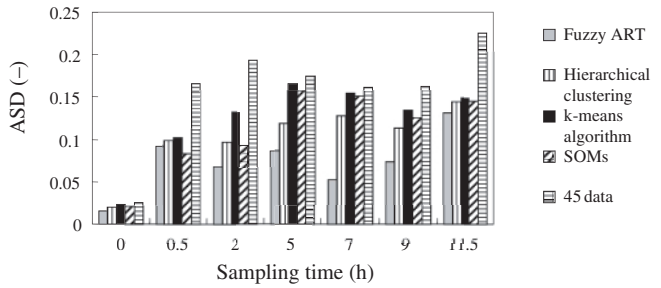

Fig. 7. Comparison of ASD.

that successful clusterings for $t = 2, 5, 7, 9$ are achieved only by Fuzzy ART.

Comparison of clustering robustness against noise

In order to compare the clustering repeatability, we generated five sets of randomly noised data. Generally, the fluctuation of microarray data is within about 2-fold change, and we added a random value from -1.0 to 1.0 to the $\log_2 R_t$ value. Table 6 shows the results of clustering robustness using five sets of noised data in four clustering methods. In the case of Fuzzy ART, total 178 genes among 225 (45 genes $\times$ 5 sets) genes were clustered into the same clusters as those using un-noised data. That is to say, 79.1% of the genes were preserved in terms of robustness after adding random noise. We defined the robustness ratio as the ratio of genes whose clustering result was coherent. In the cases of hierarchical clustering, k-means algorithm and SOMs, robustness ratios were 73.3, 55.6 and 57.3%, respectively. It is obvious that the clustering result by Fuzzy ART is the highest score. The correctness ratios defined above were also calculated for 5 sets with the noised data. In the case of Fuzzy ART, total 89 genes among 105 (21 index genes $\times$5 sets) genes were clustered correctly, and the correctness ratio was 0.85. In the cases of the hierarchical clustering, k-means algorithm and SOMs, the correctness ratios were 0.78, 0.80 and 0.82, respectively. It is also clear that the Fuzzy ART achieved the highest correctness ratio. From these results, it was shown that Fuzzy ART is also more useful than the other clustering methods in the case of noised data.

Biological validation of clustering result by Fuzzy ART

In Chu and Herskowitz (1998), expression levels of *DMC1*, *NDT80*, *SPS1* and *DIT1* by the northern analysis were shown. It was shown that *DMC1* gene expressed during relatively early phase of sporulation. This fact agrees well with the characterization of the cluster 1 by Fuzzy ART. It was also shown that both *NDT80* and *SPS1* expressed during 5–9 hours after transfer to SPM. It is also reasonable that these two genes were clustered in the same cluster 5 in Fuzzy ART. Furthermore, *DIT1* gene expressed temporally and specifically during 7–9 hours after transfer to SPM. This gene was correctly clustered in the distinct 'Mid–Late' cluster 4 in Fuzzy ART.

In Sym *et al.* (1993), the expression level of *ZIP1* gene was monitored using β-galactosidase assay. The activity of β-galactosidase increased gradually, peaked about 6 hours after sporulation and gradually decreased. This profile of β-galactosidase assay corresponded to the weight vector of the cluster 2, and it is reasonable that *ZIP1* gene was grouped in the cluster 2 in Fuzzy ART. From the comparison of the other clustering results for *DMC1*, *NDT80*, *SPS1* and *ZIP1* genes, it is said that the result of hierarchical clustering is different from those of the k-means algorithm, SOMs and Fuzzy ART. As discussed above, Fuzzy ART could only classify *DIT1* gene as 'Mid–Late' cluster among the four methods.

Finally, we discuss *SPS100* gene, which is not included in Table 1. *SPS100* gene was not selected through the data preprocessing in this study, since we followed the criteria mentioned in the theoretical analysis section of Chu *et al.* (1998). They selected genes which show 2.2-fold change during sporulation. However, *SPS100* gene was used to create an average temporal profile of 'Late' induced genes in the biological analysis section of Chu *et al.* (1998). Briza *et al.* (1990) also reported that *SPS100* gene expressed 14 hours after transfer to SPM. Therefore, we added *SPS100* gene profile to 45-gene profile, and clustered again the 46-gene profile by Fuzzy ART. Figure 8 shows the clustering result. The *SPS100* gene profile was only classed as the newly generated cluster 6 in the same condition used above. The other 45 genes were clustered in the same way mentioned above. This result means that Fuzzy ART can cluster a distinctly different profile, such as *SPS100* gene, as another cluster. Hierarchical clustering method also classified *SPS100* as another cluster, but SOMs could not. In the case of k-means algorithm, since we need to decide the number of k previously, it is impossible to discuss the clustering results in the same manner. However, k-means algorithm could

Table 6. Comparison of clustering results using noised data

	Fuzzy ART	Hierarchical clustering	k-means algorithm	SOMs
Average of clustering robustness	79.1	73.3	55.6	57.3
Average of correctness ratio	0.85	0.78	0.80	0.82

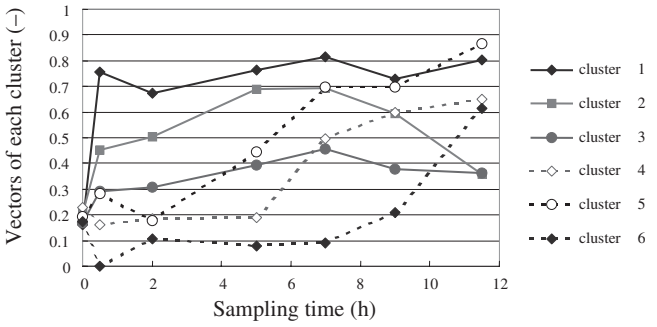

Fig. 8. Weight vector W of six clusters generated by Fuzzy ART using 46 genes' profiles including *SPS100* 'Late' gene.

not classify *SPS100* gene alone as another cluster even if we set $k = 6$.

In a future study, we expect to identify the induction time of other sporulation-specific genes using Fuzzy ART and elucidate the regulation of genes. For this aim, one of the investigations we are focusing on is the data preprocessing. *IME1*, for example, is an essential transcription factor that fundamentally regulates the initiation of yeast sporulation. Nevertheless, *IME1* was cut off since it did not show the significant change of expression ratio at any sampling point (Chu *et al.*, 1998). Therefore, the cut off preprocessing is also a very important step for the analysis of expression data. We are also investigating the application of Fuzzy ART to cluster other experimental data, especially of human cancer cells. Successive clustering of cancer cells may lead to the development of an optimal medical treatment toward those cancer cells.

CONCLUSION

In this paper, we clustered the 45 sporulation-specific genes and verified the advantage of using Fuzzy ART as a clustering method for expression data. It was found that the clustering result of Fuzzy ART only showed the successful classification of 'Mid–Late' genes, such as *DIT1* and *DIT2* at the number of clusters 5. In the mathematical validation, it is clear that the average gap index of Fuzzy ART is the lowest. Comparison based on the distribution of profiles also proved that only Fuzzy ART achieved successful clustering. We verified the robustness of Fuzzy ART with noised data. Through verifications for biological validations, clustering results by Fuzzy ART corresponded well to existing biological knowledge.

ACKNOWLEDGEMENTS

This research was supported in part by Grant-in-Aid for Scientific Research on Priority Areas (2) 'Genome Informatics Science' (No. 14015228) from the Ministry of Education, Science, Sports and Culture of Japan.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Briza,P., Breitenbach,M., Ellinger,A. and Segall,J. (1990) Isolation of two developmentally regulated genes involved in spore wall maturation in *Saccharomyces cerevisiae*. *Genes Dev.*, **4**, 1775–1789.

Carpenter,G.A. and Grossberg,S. (1987a) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vis. Grap. And Image Process.*, **37**, 54–115.

Carpenter,G.A. and Grossberg,S. (1987b) ART 2: stable self-organization of pattern recognition codes for analog input patterns. *Applied Optics*, **26**, 4919–4930.

Carpenter,G.A., Grossberg,S. and Reynolds,J.H. (1991a) ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, **4**, 565–588.

Carpenter,G.A., Grossberg,S. and Rosen,D.B. (1991b) Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks*, **4**, 759–771.

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. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W. and Lockhart,D.J. (2001) Transcriptional regulation and function during the human cell cycle. *Nature Genet.*, **27**, 48–54.

Chu,S. and Herskowitz,I. (1998) Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. *Mol. Cell*, **1**, 685–696.

Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.

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

Frank,T., Kraiss,K. and Kuhlen,T. (1998) Comparative analysis of fuzzy ART and ART-2A network clustering performance. *IEEE Tran. Neural Networks*, **9**, 544–559.

Hartigan,J.A. (1975) *Clustering algorithm*. Wiley, New York.

Khodursky,A.B., Peter,B.J., Cozzarelli,N.R., Botstein,D., Brown,P.O. and Yanafsky,C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **97**, 12170–12175.

Kupiec,M., Byers,B., Esposito,R.E. and Mitchell,A.P. (1997) The molecular and cellular biology of the yeast Saccharomyces, pp. 889–1036. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Lashkari,D.A., DeRisi,J.L., McCusker,J.H., Namath,A.F., Gentile,C., Hwang,S.Y., Brown,P.O. and Davis,R. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA*, **94**, 13057–13062.

Mitchell,A.P. (1994) Control of meiotic gene expression in *Saccharomyces cerevisiae*. *Microbiol. Rev.*, **58**, 56–70.

Perou,C.M., Jeffrey,S.S., Rijn,M.V.D., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C.F., Lashkari,D., Shalon,D., Brown,P.O. and Botstein,D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Ress,C., Spellman,P., Iyer,V., Jeffrey,S.S., Rijn,M.V.D., Waltham,M., Pergamenschikov,A., Lee,J.C.F., Lashkari,D., Shalon,D., Myers,T.G., Weinstein,J.N., Botstein,D. and Brown,P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–235.

Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T., Scudiero,D.A., Eisen,M.B., Sausville,E.A., Pommier,Y., Botstein,D., Brown,P.O. and Weinstein,J.N. (2000) A gene expression database for the molecular pharmacology of cancer. *Nature Genet.*, **24**, 236–245.

Somogyi,R. (1999) Making sense of gene-expression data. *Pharmainformatics*, 17–24.

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.

Sym,M., Engebrecht,J. and Roeder,S. (1993) ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis. *Cell*, **72**, 365–378.

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 with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.