Membership Scoring via Independent Feature Subspace Analysis for Grouping Co-Expressed Genes

Heyjin Kim, Seungjin Choi, Sung-Yang Bang
Dept. of Computer Science and Engineering
POSTECH
San 31 Hyoja-dong, Nam-gu
Pohang 790-784, Korea
Email: {marisan, seungjin, sybang}@postech.ac.kr

***Abstract*— Linear decomposition models such as principal component analysis (PCA) and independent component analysis (ICA) were shown to be useful in analyzing high dimensional DNA microarray data, compared to clustering methods. Assuming that gene expression is controlled by a linear combination of uncorrelated/indepdendent latent variables, linear modes were shown to be related to some biological functions. However, grouping co-expressed genes using these methods is not quite successful since they take some biological dependence into account. In this paper, we employ the independent feature subspace analysis (IFSA) method [8] which finds phase- and shift-invariant features. We propose a new membership scoring method based on invariant features from IFSA and show its usefulness in grouping functionally-related genes in the presence of time-shift and expression phase variance. This is confirmed through *PathCalling*.**

I. Introduction

Current DNA microarray technology produces a huge amount of high dimensional data and enables us to measure the expression levels of thousands of genes simultaneously. One interesting task in microarray data analysis is to monitor gene expression levels while a cell undergoes some biological process, in order to find co-expressed genes. One of exemplary data might be the yeast *Saccharomyces cerevisiae* whose genome sequences were revealed and the Open Reading Frames (ORFs) were already determined. Chu *et al.* [4] and Spellman *et al.* [14] studied cell cycle behavior using the microarray technique to analyze the roles of specific genes in the process.

A variety of methods have been employed for the analysis and interpretation of gene expression data. These include: (1) clustering methods such as hierarchical clustering and self-organizing map; (2) linear decomposition models such as PCA and ICA. Recently it was shown that linear models-based methods (especially ICA-based methods) are useful in DNA microarray data analysis [7], [10], [11], [13]. Linear models describe the expression levels of genes as linear functions of latent variables (hidden variables) which might be related to distinct biological causes of variation such as regulators of gene expression, cellular functions, or responses to experimental treatments [10]. Martoglio *et al.* [11] suggested to use a linear model in which each gene can participate, to varying degrees, in many independent patterns of covariation since each gene is expected to be influenced by several transcription factors, each of which influences several genes.

Grouping co-expressed genes is an important task [5] because co-expressed genes help us to explore regulatory networks, to find transcription factor binding sites, or to discover a certain biological function of a gene. Linear model methods find linear modes and influences. These linear modes are served as prototype patterns in order to find co-expressed genes. This is carried out through pattern matching. However a simple pattern matching is not always successful in grouping co-expressed genes. In a co-regulated system, some genes that are functionally related, do not share similar patterns and genes which do not have functional relations, show similar patterns. This is why grouping co-expressed genes is a difficult problem.

In this paper we take a close look at some biological dependency that linear model-based methods (such as PCA, ICA, and Bayes Decomposition) [10]–[13] overlooked. To this end, we employ the independent feature subspace analysis (IFSA) that was originally developed by Hyvärinen and Hoyer [8] in order to study a computational model for a complex cell in V1 (primary visual cortex). Since IFSA finds phase- and shift-invariant features, it is expected to grouping slightly time-delayed patterns as well as exactly matched pattern. Here we propose a new method for grouping co-expressed genes, which is based on membership scoring based on invariant features from IFSA. The usefulness of our method is verified by PathCalling and experimental evidence in [3], [6], [16].

II. Methods

A. Independent Feature Subspace Analysis

Linear decomposition model assumes that the data matrix $\boldsymbol{X} = [x_{ij}]$ (where the element x_{ij} represents the expression level of gene i in the jth sample, $i = 1, \ldots, m$, $j = 1, \ldots, N$) is modelled as

$$\boldsymbol{X} = \boldsymbol{S}\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a matrix consisting of latent variables and the row vectors of $\boldsymbol{A} \in \mathbb{R}^{n \times N}$ are basis vectors (corresponding to *linear modes* in [10]).

ICA aims at finding a representation (1) with latent variables (contained in the columns of $\boldsymbol{S}$) being statistically independent. The statistical independence among latent variables is a key assumption (as well as a limitation) in ICA. Multidimensional ICA [2] generalized the ICA by allowing

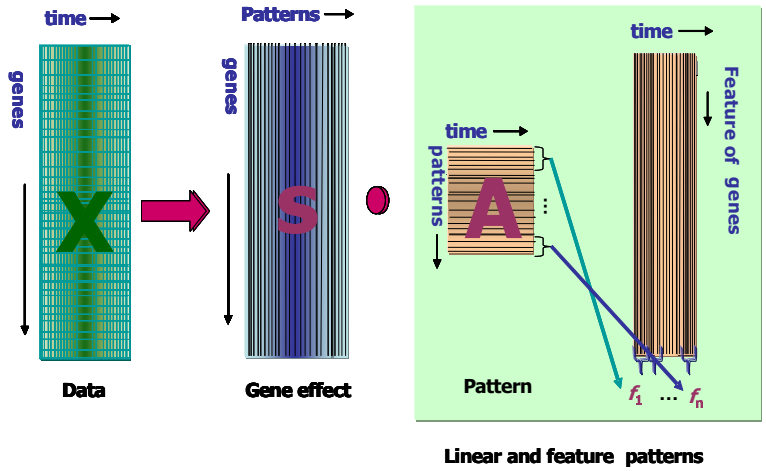


Fig. 1. The microarray data matrix $\boldsymbol{X}$ is decomposed into a product of two matrices $\boldsymbol{S}$ (gene effect) and $\boldsymbol{A}$ (linear modes). Invariant feature patterns $\{f_j\}_{j=1}^{J}$ (computed by (2)) play a critical role in grouping similar patterns with preserving shift- and phase-invariance.

the components in a p-tuple to be dependent but requiring different p-tuples to be independent. The IFSA [8] embedded the invariant feature subspaces in multidimensional ICA by considering the probability distributions for the p-tuples of latent variables that are spherically symmetric, i.e., depend only on the norm (see Fig. 1 for pictorial elucidation of IFSA as well as linear decomposition model). In contrast to ICA, the IFSA aims at finding a linear transformation $\boldsymbol{W}$ (which corresponds to the inverse system of $\boldsymbol{A}$) such that feature subspaces (obtained by taking square root of sum of energy of responses) become independent but components in a feature subspace is allowed to be dependent.

We assume that the data matrix $\boldsymbol{X}$ is already whitened. In other words, the row vectors of $\boldsymbol{A}$ are confined to be orthogonal each other and to be normalized to have unit norm. Non-orthogonal factor is reflected in a whitening transform. In order to avoid an abuse of notations, we use the notation $\boldsymbol{X}$ for the whitened data matrix.

Let's denote by $\boldsymbol{x}_i$, the ith row vector of $\boldsymbol{X}$ and by $\boldsymbol{w}_i$, the ith column vector of $\boldsymbol{W}$ (which is the inverse system of $\boldsymbol{A}$). Note that $\boldsymbol{w}_i$ corresponds to the ith row vector of $\boldsymbol{A}$ since orthogonal basis vectors are considered here. We consider the case where latent variables are divided into J number of p-tuples (where p represents the dimension of subspace). For the sake of simplicity, we assume identical dimension, p for every feature subspace. The jth feature subspace is denoted by $\mathcal{F}_j$. The value $f_j(\boldsymbol{x})$ in $\mathcal{F}_j$ with data vector $\boldsymbol{x}$ is given by

$$f_j(\boldsymbol{x}) = \sum_{i \in \mathcal{F}_j} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle^2 , \tag{2}$$

where $< \cdot, \cdot >$ is the inner product. In fact $f_j(\boldsymbol{x})$ is a pooled energy.

With these notations, we can write the log-likelihood $\mathcal{L}$ of the data given the model as

$$\begin{aligned} \mathcal{L} &= \sum_{t=1}^{m} \sum_{j=1}^{J} \log p \left(\sum_{i \in \mathcal{F}_j} \langle \boldsymbol{w}_i, \boldsymbol{x}_t \rangle^2 \right) \\ &\quad + m \log |\det \boldsymbol{W}| , \end{aligned} \tag{3}$$

where $p\left(\sum_{i \in \mathcal{F}_j} s_i^2\right) = p_j(s_i, i \in \mathcal{F}_j)$ $(s_i = \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle)$ represents the probability density inside the jth p-tuple of s_i.

The IFSA finds a linear transform $\boldsymbol{W}$ which maximizes the log-likelihood (3). Learning independent feature subspaces is carried out by a stochastic gradient ascent method, whose updating rule has the form

$$\Delta \boldsymbol{w}_i \propto \boldsymbol{x} \langle \boldsymbol{w}_i, \boldsymbol{x} \rangle \varphi \left(\sum_{r \in \mathcal{F}_{j(i)}} \langle \boldsymbol{w}_r, \boldsymbol{x} \rangle^2 \right) , \tag{4}$$

where $j(i)$ is the index of the feature subspace which $\boldsymbol{w}_i$ belongs to and φ is the score function, i.e., $\varphi = \frac{p'}{p}$. More details on IFSA (including the description of hypothesize density p) can be found in [8].

B. Membership Scoring for Clustering

Linear model-based methods (such as Bayes Decomposition, PCA, and ICA) showed that linear modes (corresponding to basis vectors) might be related with certain biological functions such as phases of cell cycle or mating response, etc. [10], [12], [13]. For instance, Liebermeister [10] showed that linear modes that are estimated by ICA match B-cell activation and Lymph node, and so on. However these methods do not take some dependency among biological patterns into account. Here we focus on some connections between related functions with close time intervals and made a group of functionally linked patterns.

In contrast to ICA-based methods where only the influence of gene ($\boldsymbol{S}$) was exploited for grouping similar patterns, we introduce a method of membership scoring which is described in details below. We normalize each row vector of the gene influence matrix $\boldsymbol{S}$ using l_1 norm and denote the resulting matrix by $\widetilde{\boldsymbol{S}} = [\tilde{s}_{ij}]$. In other words, each row vector in $\widetilde{\boldsymbol{S}}$ has unit l_1 norm. The feature value (which results from spatial pooling) $f_{ik} = \sum_{l \in \mathcal{F}_k} \langle \boldsymbol{w}_l, \boldsymbol{x}_i \rangle^2$ is also normalized over k. This normalized feature values are denoted by $\tilde{f}_{ik}$. With these normalized gene influences and feature values, we compute a membership scoring values ψ_{ik} by

$$\psi_{ik} = \tilde{f}_{ik} \left| \tilde{s}_{ij} \right|, \quad k = 1, \ldots, J \tag{5}$$

For each gene, J possible membership scoring values, $\{\psi_{i1}, \ldots, \psi_{iJ}\}$ are computed. Then, a maximal membership scoring, ψ_i^{max} is calculated over k, i.e.,

$$\psi_i^{max} = \max_k \psi_{ik}, \tag{6}$$

This maximal membership scoring ψ_i^{max} indicates an appropriate cluster that the ith gene is expected to belong to. In order to filter out some irrelevant genes, we threshold out genes which have relatively very low maximal membership scoring values. The level of threshold is decided, depending on how many relevant genes we want to group together.

Our results (see Sec. III) confirm that this new membership scoring method well explain the relationship between linked genes (that were already revealed by many experiments in biology domain)

III. Results

We apply our membership scoring method based on IFSA to the yeast cell cycle data from Spellman *et al.* [14], which contains the expression of 6178 ORFs during the cell replication cycle in the budding yeast *Saccaromyces cerevisiae*. This data set contains 77 tissue samples in different experimental conditions such as α factor pheromone, cdc15, cdc28, elucidation, and so on.

Throughout several subsets of data, we confirm that our method is quite useful in grouping co-expressed genes. We also compared our results to PCA- or ICA-based methods, but due to the space limitation we do not include comparison results. These comparison results can be found in our supplement web site [1]. In grouping co-expressed genes, our method was quite successful, whereas PCA- or ICA-based methods were not. For fair comparison, we pre-processed the data by time warping [1], [9] and spline interpolation before we apply our method as well as ICA.

One of useful property in IFSA is that it finds a shift- and phase-invariant features. In gene expression time series, this invariant property produces a set of profiles containing slightly time-delayed patterns or some contrary patterns that show direction opposite to a main direction in a group. This invariance property is a key ingredient to our successful grouping, compared to previous linear model-based methods. In order to take this invariance property into account, we include slightly time-delayed patterns or some contrary patterns as well as exactly matched patterns in a group of co-expressed patterns. In other words, we let IFSA to find a statistical structure of gene expression data in an unsupervised fashion and rely on this result for grouping co-expressed genes. It turns out that this approach is able to consider some biological dependence in a task of grouping, whereas ICA-based methods do not take into account.

In general, it is not easy to interpret functional relations among several tens of genes. Hence we take a small size of data so that the interpretation become easier. We use 4 different test data sets (which is summarized in Table I). For graphical representation for gene interactions in test data sets in Table I, we used a web interface software, *PathCalling* [16], which provides information on putative protein interactions identified in the screens or reported in the literature. The output of PathCalling is remodelled in a form of undirected graph (see Fig. 2). Each node represents a gene and edges offer insight into novel interactions between proteins involved in the identical biological function.

Experiment 1: In the first experiment, we used the Test Data Set 1 and decomposed it into two groups (i.e., $J = 2$) with each group having two components ($p = 2$) by IFSA. The result is summarized in Table II. Every gene associated with the 1st component in Group 1 has its peak at G1 phase. Functions in group 1 mainly focus on DNA replication and G1/S cyclin. On the other hand, the genes in Group 2 are involved with G2/M cyclin. In this experiment we used a threshold value in the range between 10^{-4} and 10^{-5}. This threshold value is used in the rest of experiments.

Experiment 2: In this experiment, we investigated the effect of an extraneous gene YLR190W (which does not have any functional relation with other genes in the data set) whether our method is influenced by this artifact-like gene or not. The result summarized in Table III shows this unrelated gene is excluded, i.e., its maximal membership score value is below the threshold. The Data Set 2 was decomposed into three groups with dimension 2 (i.e., $J = 3,\ p = 2$). Genes associated with the 1st component in Group 1 are Pho85 cyclin family (PCL family) which activates Pho85 as a CDK [15]. PCL1, PCL2 and PCL9 are the only members expressed in a cell cycle-regulated pattern; PCL9 is activated by the transcription factor SWI5 and PCL1/PCL2 by SBF. One interesting point in this experiment is in PHO85. The gene PHO85 plays an important role in Group 3, the analysis method based on gene phase pattern (which is quite common method in the study of cell cycles) has difficulty in grouping this gene from only phase).

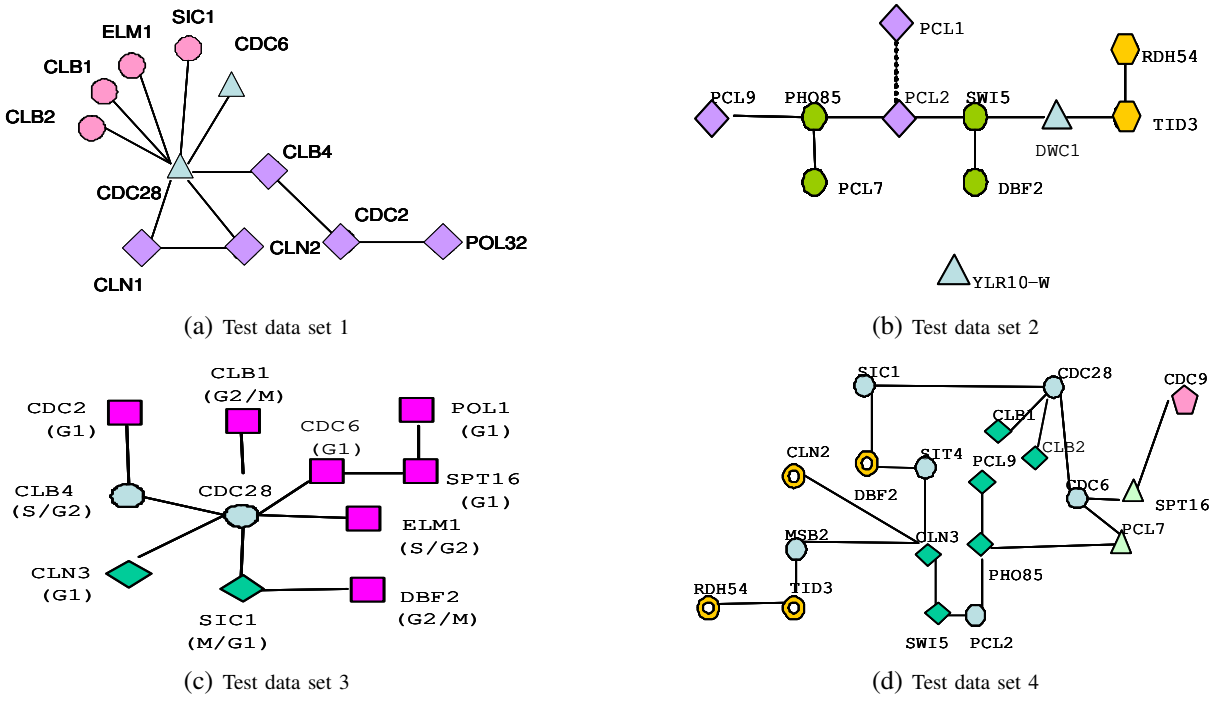
Experiment 3: In this experiment, we used Test Data Set 3 which contains all the genes in Test Data 1 and some other genes (see Table IV). Like the Experiment 1, we decomposed the data into 2 groups with dimension 2. The result is summarized in Table IV. We observed that our proposed

[1] http://home.postech.ac.kr/∼marisan/Bioinformatics/ISAcycle.htm

TABLE I

DATA SETS USED IN EXPERIMENT 1-4. EACH DATA SET CONTAINS A SET OF GENES WHICH HAVE SOME INTERACTIONS. ONLY ONE GENE YLR190W (UNDERLINED) DOES NOT HAVE ANY RELATION WITH OTHER GENES. THESE DATA SETS ARE USED TO TEST THE VALIDITY OF OUR METHOD.

Data Set	Set I	Set II	Set III	Set IV
Gene Names (SGD)	CDC2 CDC6 CLB1 CLB2 CLB4 CLN1 CLN2 ELM1 POL32 SIC1	PCL9 SWI5 PCL2 DBF2 RDH54 TID3 PCL7 PHO85 PCL1 YLR190W	CDC2 CDC5 CDC6 CDC9 CLB1 CLB2 CLB4 CLN1 CLN2 CLN3 DBF2 ELM1 HYS2 POL1 POL32 SIC1 SPT16	CDC2 CDC5 CDC6 CDC9 CLB1 CLB2 CLB4 CLN1 CLN2 CLN3 DBF2 ELM1 HYS2 PCL1 PCL2 PCL7 PCL9 PHO85 POL1 POL32 RDH54 SIC1 SPT16 SWI5 TID3 YLR190W



(a) Test data set 1

(b) Test data set 2

(c) Test data set 3

(d) Test data set 4

Fig. 2. Undirected graphical representation for gene interaction in each test data set summarized in Table I: (a) In test data set 1, circle and diamond indicate two different groups, mainly CLB families and CLN families, respectively. Triangle shape corresponds to genes that do not belong to anywhere.; (b) In addition to circle, diamond, and triangle (which represent the same meaning as those in (a)), two hexagons indicate another group, most of which are PCL families and their regulatory genes; (c) Square and diamond indicate two different groups and circle-shaped nodes represent genes that do not belong to these two groups. The erroneous gene DBF2 connected with CLN3 and SIC1 (a protein inhibitor), belong to another group, members of which function as a protein kinase; (d) Nodes with different shapes (except for circle-shaped nodes) represent different groups. The circle-shaped nodes indicate the genes which do not belong to any group. In Experiment 4, some genes belonging to the same group were successfully grouped by our method.

method were able to identify some contrary profiles that are usually considered as different patterns in existing methods. Expression profiles associated with genes POL1, CDC2, and CDC6, are in the opposite directions to those of ELM1, CLB1 and DBF2 (see Fig. the expression profiles for CLN3 and SIC 1 which were known to have similar patterns and to have peaks at G2/M and M/G1, respectively.

Experiment 4: The Test Data Set 4 contains genes in Test Data Set 1-3. Although we increased the data size, grouping based on our method was preserved. See the result in Table V. For instance, genes SWI5 and PHO85 were grouped together both in Experiment 2 and in Experiment 4.

IV. DISCUSSION

In this paper, we proposed a new method of grouping co-expressed genes using the membership scoring based on IFSA. In contrast to ICA-based methods, IFSA employed a linear decomposition with spatial pooling so that invariant features could be obtained. The invariant features (which are pooled energy) was nicely incorporated into our membership scoring. Our simulation results showed that the proposed method well matched the relationship between linked genes that were already known in biology community.

TABLE II

RESULT IN EXPERIMENT 1 : CLB FAMILY, CLN FAMILY, AND SOME RELATED GENES.

Group	peak	ψ_i^{max}	***SGD***	Process	Function
1-1	G1	0.02239	CDC2	DNA replication	DNA polymerase δ catalytic 125kd subunit
	G1	0.03510	POL32	DNA replication	polymerase δ 55kd subunit
	G1	0.01233	CLN2	cell cycle	G1/S cyclin
1-2	G1	0.00195	CLN1	cell cycle	G1/S cyclin
2-1	S/G2	0.12880	ELM1	growth	protein kinase
	G2/M	0.00187	CLB1	cell cycle	G2/M cyclin
	G2/M	0.01579	CLB2	cell cycle	G2/M cyclin
	M/G1	0.07794	SIC1	cell cycle	Cdc28p-Clb5 protein kinase inhibitor

TABLE III

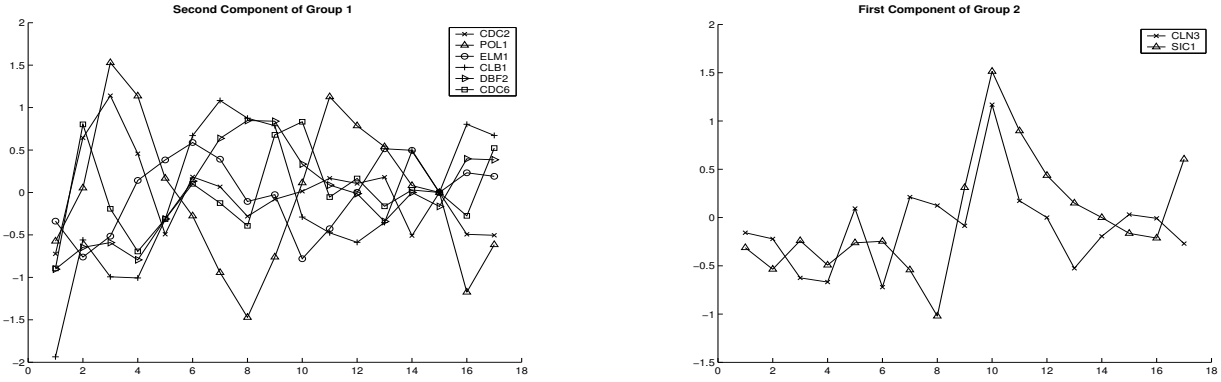
RESULT IN EXPERIMENT 2: EXTRANEOUS GENE, YLR10W AND PCL FAMILY.

Group	peak	ψ_i^{max}	***SGD***	Process	Function
1-1	G1	0.01790	PCL2	cell cycle	G1/S cyclin
	G1	0.00236	PCL1	cell cycle	G1/S cyclin
	M/G1	0.00152	PCL9	cell cycle	cyclin (Pho85p)
2-2	G1	0.00407	RDH54	meiosis	helicase
	S	0.00272	TID3	unknown	unknown
3-2	S/G2	0.01063	PCL7	cell cycle	cyclin
	G2/M	0.00060	SWI5	cell cycle	transcription factor, regulates HO
	G2/M	0.00212	DBF2	cell cycle	late mitosis; protein kinase
	U	0.00349	PHO85	cell cycle	cyclin-dependent protein kinase

TABLE IV

RESULT IN EXPERIMENT 3: PROTEIN KINASE VERSUS PROTEIN KINASE INHIBITOR.

Group	peak	ψ_i^{max}	***SGD***	Process	Function
1-1	G1	0.00057	SPT16	chromatin structure	non-histone protein
1-2	G1	0.21248	POL1	DNA replication	polymerase α 180kd subunit
	G1	0.00068	CDC2	DNA replication	polymerase δ 125kd subunit
	M/G1	0.00188	CDC6	DNA replication	pre-initiation complex formation
	S/G2	0.06263	ELM1	pseudohyphal growth	protein kinase
	G2/M	0.00116	CLB1	cell cycle	G2/M cyclin
	G2/M	0.01169	DBF2	cell cycle	late mitosis; protein kinase
2-1	G2/M	0.07817	CLN3	cell cycle	G1/S cyclin
	M/G1	0.03980	SIC1	cell cycle	Cdc28p-Clb5 protein kinase inhibitor



(a) Expression profiles which correspond to the second component in group 1. (b) Expression profiles which correspond to the first component in group 2.

Fig. 3. Expression Profiles in Experiment 3.

TABLE V
RESULT IN EXPERIMENT 4

Group	peak	ψ_i^{max}	***SGD***	Process	Function
2-2	G1	0.00240	SPT16	chromatin structure	non-histone protein
	S/G2	0.00278	PCL7	cell cycle	cyclin
3-1	G1	0.00146	CLN2	cell cycle	G1/S cyclin
	G1	0.00056	RDH54	meiosis	helicase
	S	0.00467	TID3	unknown	unknown
	G2/M	0.00190	DBF2	cell cycle	late mitosis, protein kinase
4-1	G1	0.00655	CDC9	DNA replication/repair	DNA ligase
6-1	G2/M	0.00181	CLN3	cell cycle	G1/S cyclin
	G2/M	0.00108	SWI5	G1 cell cycle	transcription factor, regulates HO
	U	0.00857	PHO85	cell cycle	cyclin-dependent protein kinase
6-2	G2/M	0.00087	CLB1	cell cycle	G2/M cyclin
	G2/M	0.00096	CLB2	cell cycle	G2/M cyclin
	M/G1	0.01411	PCL9	cell cycle	cyclin (Pho85p)

ACKNOWLEDGMENT

Authors would like to thank A. Hyvärinen and P. Hoyer for sharing their IFSA code. Portion of this work was carried out when S. Choi visited Advanced Brain Signal Processing Lab in RIKEN, Japan. This work was supported by Basic Science Research Center, POSTECH, by Korea Ministry of Science and Technology under Brain Science and Engineering Research Program and an International Cooperative Research Project, and by Ministry of Education of Korea for its financial support toward the Electrical and Computer Engineering Division at POSTECH through its BK21 program.

REFERENCES

[1] J. Aach and G. M. Church, "Aligning gene expression time series with time warping algorithms," *Bioinformatics*, vol. 17, no. 6, pp. 495–508, 2001.

[2] J. F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP*, Seattle, WA, 1998.

[3] J. M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer, and D. Botstein, "Genetic and physicalmaps of *saccharomyces cerevisiae*," *Nature*, vol. 387, pp. 67–73, May 1997.

[4] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lcokhart, and R. W. Davis, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, vol. 2, pp. 65–73, 1998.

[5] L. J. Heyer, S. Kruglyak, and S. Yooseph, "Exploring expression data: Identification and analysis of coexpressed genes," *Genome Res.*, vol. 9, pp. 1106–1115, 1999.

[6] Y. Ho, A. Gruhler, A. Hellbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutillier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Serensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers, "Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry," *Nature*, vol. 315, no. 10, pp. 180–183, Jan. 2002.

[7] G. Hori, M. Inoue, S. Nishimura, and H. Nakahara, "Blind gene classification based on ICA of microarray data," in *Proc. ICA*, San Diego, California, 2001.

[8] A. Hyvärinen and P. O. Hoyer, "Emergence of phase and shift invarianct features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[9] E. Keogh and M. Pazzani, "Derivative dynamic time warping," in *Proc. SIAM Int. Conf. Data Mining*, 2001.

[10] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, 2002.

[11] A. M. Martoglio, J. W. Minskin, S. K. Smith, and D. J. C. MacKay, "A decomposition model to track gene expression signatures: Preview on observer-independent classification of ovarian cancer," *Bioinformatics*, vol. 18, no. 12, pp. 1617–1624, 2002.

[12] T. D. Moloshok, R. R. Klevecz, J. D. Grant, F. J. Manion, W. F. S. IV, and M. F. Ochs, "Application of bayesian decomposition for analysing microarray data," *Bioinformatics*, vol. 18, no. 4, pp. 566–575, 2002.

[13] S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal components analysis to summarize microarray experiments: Application to sporulation time series," in *Proc. Pacific Symp. Biocomputing*, 2000, pp. 452–463.

[14] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273–3297, Dec. 1998.

[15] C. N. Tennyson, J. Lee, and B. J. Andrews, "A role for the Pcl9-Pho85 cyclin-cdk complex at the M/G_1 boundary in *saccharomyces cerevisiae*," *Molecular Microbiology*, vol. 28, no. 1, pp. 69–79, 1998.

[16] P. Uetz, L. Gilot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and M. Rothberg, "A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*," *Nature*, vol. 403, pp. 623–631, Feb. 2000.