BLIND GENE CLASSIFICATION BASED ON ICA OF MICROARRAY DATA

Gen Hori[1], Masato Inoue[2,3], Shin-ichi Nishimura[2,4] and Hiroyuki Nakahara[2]

1 Laboratory for Advanced Brain Signal Processing,
Brain Science Institute, RIKEN, Saitama 351-0198, Japan

2 Laboratory for Mathematical Neuroscience,
Brain Science Institute, RIKEN, Saitama 351-0198, Japan

3 Department of Otolaryngology, Graduate School of Medicine,
Kyoto University, Kyoto 606-8507, Japan

4 Department of Otolaryngology, Faculty of Medicine,
University of Tokyo, Tokyo 113-8656, Japan

mailto:fhori@bsp.,minoue@,nishi@mns.,hiro@gbrain.riken.go.jp

ABSTRACT

The present study shows that an ICA-based method can, effectively and blindly, classify a vast amount of gene expression data into biologically meaningful groups. Specifically, we show (1) that genes, whose expression data are sampled at different times, can be classified into several groups, based on the correlation of each gene with independent component curves over time, and (2) that these classified groups by ICA-based method have a good match with the classified groups that are determined by use of domain knowledge and considered to be a benchmark. These results suggest that the ICA-based method can be a powerful approach to discover unknown gene functions.

1. INTRODUCTION

Recent technical advances have led to a vast amount of gene expression data, so that we have an urgent need to develop technical tools to meet this rapid expansion of data. One of important questions with given such a data is how different sets of genes work together, in other words, how we can classify genes into biologically meaningful groups. Using a specific type of data (explained below), the present study shows that ICA-based method is a promising approach to address this question.

We used a microarray data, or a gene expression data, of yeast during sporulation (process of germination), which is collected by Chu et al[3] and available in public[1]. The data consists of expression data of 6118 genes in yeast genome, which were sampled at seven different times during sporulation (namely, 0.0, 0.5, 2.0, 5.0, 7.0, 9.0 and 11.5 hours). In other words, the data can be understood as the matrix of 6118 rows and 7 columns, which correspond to the number of genes and the sampled times, respectively. Each entry in the matrix has a real value.

It is known by experiments that, during sporulation, some specific genes work at different time periods and that the expression of such genes changes significantly (positively or negatively) in comparison with the ordinary level. It is not known, however, which of many other genes work or not at different time periods during sporulation, whereas the gene expression data indicates various changes in such genes at different time periods. Some of these unknown genes may play a crucial role in sporulation. Hence, previously, several approaches are employed, including gene clustering[5], principal component analysis[6] and self-organization map[7].In the present study, we examine ICA-based method and show its validity, particularly in comparison with the method of using domain knowledge and the method of PCA.

The following section is organized as follows. Section 2 shows how the genes in the data are classified into several groups by the previous study, using domain knowledge. Section 3 shows how the PCA-based method classifies genes into several groups. In Section 4, we briefly discuss how ICA is applied to the data in the present study. We show the results in Section 5.

[1] http://cmgm.stanford.edu/pbrown/sporulation/

Discussion follows in Section 6.

2. CLASSIFICATION BY DOMAIN KNOWLEDGE

Chu et al[3], who thankfully made the data available in public, has tried to classify genes in their data into several groups, using their domain knowledge. To group the yeast genes according to their sequential induction patterns during sporulation, Chu et al hand-picked seven small sets of genes (Table 1), which are representatives of induction patterns in respective time period (which are known by previous studies, i.e., based on domain knowledge). By averaging expressions (i.e. real-values) of these genes at different times in each set, they defined seven model induction patterns over time (Fig.1). Given extensive experimental studies on yeast genes, we consider their model induction patterns as a benchmark, or as `correct' patterns.

Using this model induction pattern, all the other yeast genes are classified into one of the seven groups. Each gene is assigned to a group which shows the highest correlation coefficient between its model induction pattern and the gene expression data over time. In addition, genes within each group is ordered by the magnitude of the correlation coefficient.

Model profiles	Representative genes
Metabolic	ACS1, PYC1, SIP4 CAT2, YOR100C, CAR1
Early I	ZIP1, YDR374C, DMC1 HOP1, IME2
Early II	KGD2, AGA2, YPT32 MRD1, SPO16, NAB4 YPR192W
Early-Mid	YBL078C, QRI1, PDS1 APC4, KNR4, STU2 YNL013C, EXO1
Middle	YSW1, SPR28, SPS2 YLR227C, ORC3, YLL005C YLL012W
Mid-Late	CDC27, DIT2, DIT1
Late	SPS100, YKL050C, YMR322C YOR391C

Table 1. Yeast genes used to define average model profiles (Chu et al[3])

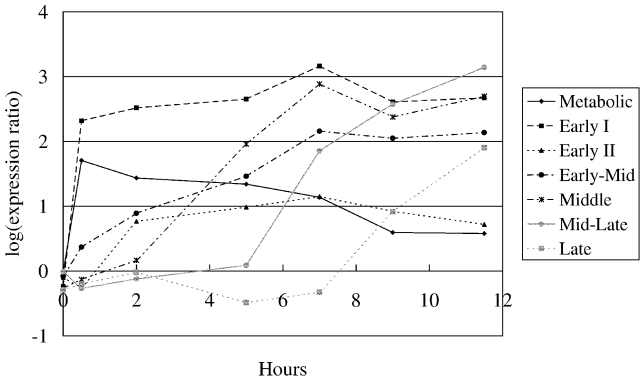

Fig.1. Model profiles obtained by averaging profiles of representative genes

3. CLASSIFICATION BY PRINCIPAL COMPONENT ANALYSIS

We first examine classification by principal component analysis (PCA), using the data provided by Chu et al[3]. PCA is in general a popular tool in analysis of gene expression data and hence, we chose to compare the classification by PCA with the classification by ICA, which is shown in the next section. In this section, we first define the classification by PCA and then shows its results. Since Raychaudhuri et al[6] applied PCA to the same data (although, for different purposes), this section partially use their results.

Let us first denote the matrix of the yeast gene expression data with rows of 6118 genes and columns of 7 conditions by X, that is, its (i,j)-th element x_{ij} represents the expression of i-th gene under j-th condition. Then the eigenvectors of the 7×7 covariance matrix of X define the principal components of the gene expression data. Let us suppose that the covariance matrix is diagonalized as $\Lambda = V^t \mathrm{Cov}(X) V$ where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_7)$, then the columns of V are the principal components. Table 2 shows all the eigenvalues λ_i $(i = 1, \ldots, 7)$ and their ratios. Fig.2 shows the first three principal components PC_i $(i = 1,2,3)$ corresponding to the largest three eigenvalues λ_i $(i = 1,2,3)$.

By examining ratios in Table 2, Raychaudhuri et al[6] observed that the first two PCs account for over 90% of the total variability and the first three do for almost 95%, and concluded that the gene expression data can be summarized in two or three variables. Also they provided the interpretations of each PC: The first component, PC_1 is almost proportional to the variance over all genes at each sampled time and distinguished genes by the overall expression levels. The PC_2 increases

almost linearly over time and distinguishes genes by their first derivatives. The third component PC_3 has concavity, or parabolic nature.

Eigenvalues	Ratios
$\lambda_1 = 2.29$	76.9%
$\lambda_2 = 0.40$	13.5%
$\lambda_3 = 0.13$	4.4%
$\lambda_4 = 0.06$	2.0%
$\lambda_5 = 0.04$	1.4%
$\lambda_6 = 0.03$	1.0%
$\lambda_7 = 0.03$	0.8%

Table 2. Eigenvalues of covariance matrix

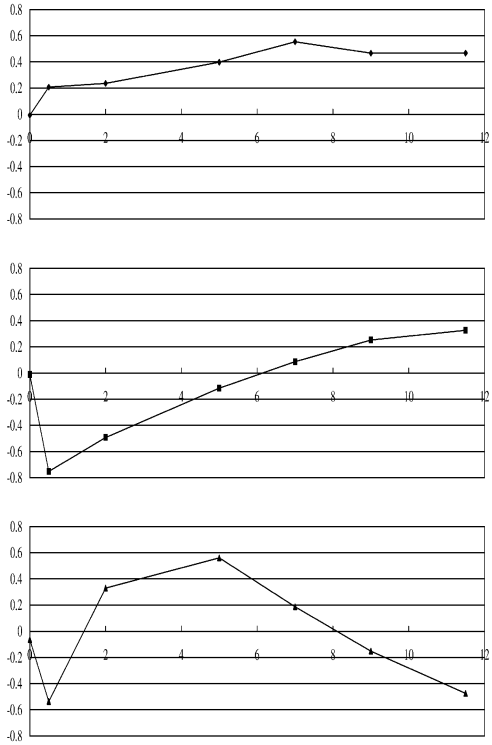

Fig.2. The first three principal components (PC_1, PC_2 and PC_3, from top to bottom) are used as model induction patterns in PCA-based gene classification.

Thus, the PCs seem to capture some features of yeast gene expression during sporulation. We can then use these PCs as model induction patterns and classify genes to groups, where the groups are defined by PCs and each gene is classified into a group of the PC that shows the highest correlation with the gene expression data over time. Genes in each group are sorted by the value of correlation coefficients. defines a new automatic gene classification method. In Section 5, we examine this results of the classification by PCA.

4. CLASSIFICATION BY INDEPENDENT COMPONENT ANALYSIS

PCA is a quite popular tool in gene expression data analysis, however, it has some limitations, simply because PCA only takes into account the second-order statistics and restricts itself to to orthogonal transformation to obtain principal components. On the other hand, independent component analysis (ICA) can take into account higher order statitics and can utilize non-orthogonal transformation for de-mixing. Hence, it is worth to examine ICA-based method for automatic gene classification.

We apply ICA to the microarray data, using the following de-mixing,

$$y = Wx;$$

where x is a 7-dimensional vector sampled from the columns of the transposed microarray data matrix X^t. We use JADE algorithm[2] to obtain the 7×7 de-mixing matrix W. Fig.3 gives the column plot of the inverse W^{-1} of the obtained de-mixing matrix. From the inverted relation $x = W^{-1}y$, one can realize that the columns of the inverse of de-mixing matrix represent how respective separated components reflect in the original microarray data. (This is similar to the columns of the inverse as \scalp map" in the case of EEG signals de-mixing.) Also the inverse of the de-mixing matrix obtained by ICA can be regarded as the counterpart of the matrix V obtained by PCA. We denote the columns of W^{-1} by $IC_i (i = 1, 2, \ldots, 7)$, for convenience.

Notably, some model induction patterns in Fig.3, blindly obtained by ICA, are similar to those in Fig.1, obtained manually depending on domain knowledge on yeast genes. For example, IC_1 and IC_2 in Fig.3 appear to match well with `Early I' and `Middle' in Fig.1, respectively.

We then use the columns of the inverse W^{-1} (or pseudo inverse) of de-mixing matrix as model induction patterns. Similarly to the previous methods, we use the correlation coefficients between a gene expression data over time and a model induction pattern and then classify each gene to the model induction pattern that shows the highest correlation. Again, genes in each group are sorted by the values of the correlation coefficients.

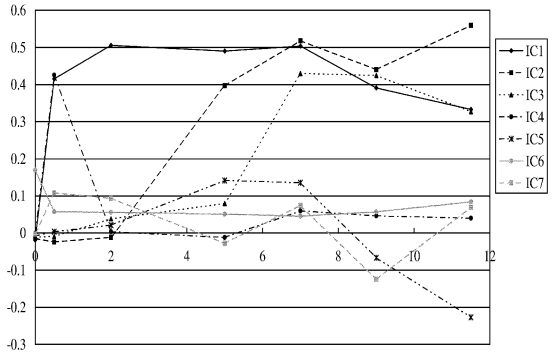

Fig.3. Columns of W $^{-1}$ sorted according to their norms in descending order

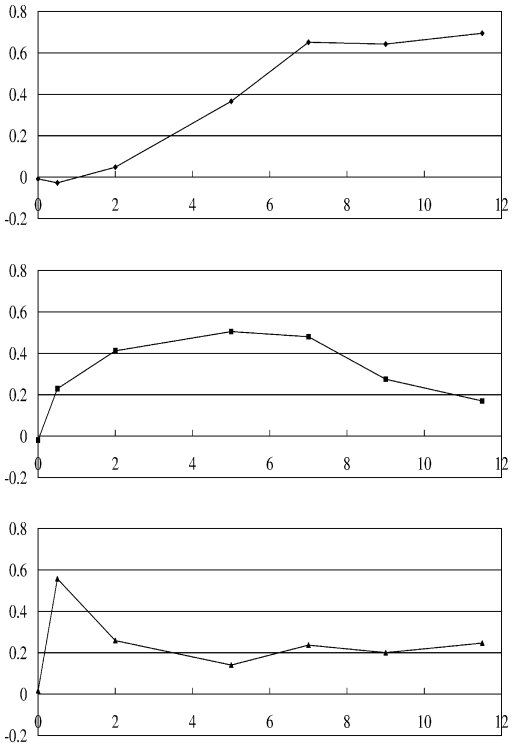

Fig.4. Columns of pseudo inverse of W (IC_1, IC_2 and IC_3, from top to bottom) are used as model induction patterns in ICA-based gene classification.

In the next section, we compare the results of the PCA-based and ICA-based gene classification methods. In this comparison, we use only top three model induction patterns for simplicity. For this purpose, we also generated three ICA-generated model induction patterns by a pre-processed cleaned data, which is done by reconstructing the data by the first three principal components. This is to eliminate ill effects from the components with small magnitudes which would not capture essential information. In this case, the demixing matrix is a 7×3 matrix and the columns of its 3×7 pseudo inverse matrix are used as model induction patterns. Fig.4 gives the column plot the pseudo inverse matrix.

5. BLIND GENE CLASSIFICATION

By results in the previous three sections, we have the following three kinds of model induction patterns: i) seven manually obtained model induction patterns by Chu et al[3] (Metabolic, Early I, Early II, Early-Mid, Middle, Mid-Late and Late), ii) three PCA-generated model induction patterns (PC_i ($i = 1;2;3$)) and iii) three ICA-generated model induction patterns (IC_i ($i = 1;2;3$)). In the following, we classify yeast genes according to the three kinds of model induction patterns and compare the ICA-generated and PCA-generated induction patterns to the manually obtained model induction patterns, which is a benchmark.

In each model induction pattern of each of the three methods, the genes are ordered by the value of their correlation coefficients. We picked the top 100 (or 200) genes in each model induction pattern and then examined how many genes of ICA method (or PCA method) overlap with the genes of model induction patterns by hand-tuned method. Table 3 and 4 show the numbers of the overlaps. (Detailed result of our gene classification is available at the author's web site [2] .)

	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3
Metabolic	0	0	0	0	0	13
Early I	13	0	0	0	7	0
Early II	2	0	0	0	0	0
Early-Mid	46	0	0	19	0	0
Middle	15	0	0	54	0	0
Mid-Late	0	6	0	7	0	0
Late	0	0	0	0	0	0

Table 3. Numbers of genes in the intersections (each group consists of 100 genes)

	PC_1	PC_2	PC_3	IC_1	IC_2	IC_3
Metabolic	0	0	0	0	20	48
Early I	46	0	0	0	48	3
Early II	24	0	6	0	19	0
Early-Mid	114	0	0	85	0	0
Middle	63	1	0	145	0	0
Mid-Late	0	31	0	48	0	0
Late	0	10	0	0	0	0

Table 4. Numbers of genes in the intersections (each group consists of 200 genes)

[2] http://www.bsp.brain.riken.go.jp/~hori/gene

Regarding the model induction patterns of the hand-picked method by Chu et al[3] as a benchmark or as `correct' patterns, we conclude that i) the classi¯ed groups by ICA-based method have a good match with the classi¯ed groups using manually obtained model induction patterns, and ii) ICA-based method classify yeast genes more distinctively than PCA-based method.

6. CONCLUDING REMARKS

The present study has shown that our ICA-based gene classi¯cation method can e®ectively classify yeast gene expressions during sporulation into biologically meaningful groups, taking the manually determined model induction patterns as a benchmark. As somewhat expected, we also showed the ICA-based method is more powerful than the PCA-based method.

Notably, our ICA-based gene classi¯cation does not require a signi¯cant amount of domain knowledge that the hand-tuned method such as the one introduced by Chu et al[3]. In other words, the present study suggests that the ICA-based gene classi¯cation may be a powerful tool for blind gene classi¯cation. Provided a rapid increase in use of microarray data, this is a strong advantage.

Let us discuss limitations and future works of the present study. First, we tested the data that contains data samples over time, i.e. the yeast gene expressions during sporulation. It is interesting to examine the validity of our method with di®erent types of gene expression data. Second, there are much to be done to combine our ICA-generated model induction patterns with other classi¯cation methods such as hierarchical gene clustering or tree harvesting. Third, in the present work, we have used correlation coe±cients between the model induction patterns and gene expression data to sort them in each group. This is just for simplicity and for easiness of comparison with other methods (i.e., the hand-tuned method by Chu et al[3]). There are other possible ways of sorting genes. For example, elements in each separated component (each row of $W X^t$) can be used as another mean of sorting genes. It may be more sensible to use other measures of similarity, like the Kullback-Leibler divergence. Fourth, rigorously speaking, raw microarray data may not be subject exactly to the linear mixing model of ICA. We should examine a more plausible assumption on the mixing, or the conditions that allows us to use linear de-mixing. Finally, we like to emphasize that the present study indicates that our ICA-based gene blind classi¯cation is promising. We need to examine the validity of the method with more data sets.

7. REFERENCES

[1] M.P.S.Brown, W.N.Grundy, D.Lin, N.Cristianini, C.W.Sugnet, T.S.Furey, M.Ares Jr. and D.Haussler, \Knowledge-based analysis of microarray gene expression data by using support vector machines", Proc. Natl. Acad. Sci. USA, 97, 1, pp.262{267, 2000.

[2] J.-F.Cardoso and A.Souloumiac, \Blind beamforming for non Gaussian signals", IEEE Proc.-F, 140, pp.362{370, 1993.

[3] S.Chu, J.DeRisi, M.Eisen, J.Mulholland, D.Botstein, P.O.Brown and I.Herskowitz, \The transcriptional program of sporulation in budding yeast", Science, 282, pp.699{705, 1998.

[4] M.B.Eisen and P.O.Brown, \DNA arrays for analysis of gene expression", Methods Enzymol, 303, pp.179{205, 1999.

[5] M.B.Eisen, P.T.Spellman, P.O.Brown and D.Botstein, \Cluster analysis and display of genome-wide expression patterns", Proc. Natl. Acad. Sci. USA, 95, 25, pp.14863{14868, 1998.

[6] S.Raychaudhuri, J.M.Stuart and R.B.Altman, \Principal component analysis to summarize microarray experiments: Application to sporulation time series", Paci¯c Symposium on Biocomputing, 5, pp.452{463, 2000.

[7] P.Tamayo, D.Slonim, J.Mesirov, Q.Zhu, S.Kitareewan, E.Dmitrovsky, E.S.Lander and T.R.Golub, \Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic di®erentiation", Proc. Natl. Acad. Sci. USA, 96, 6, pp.2907{2912, 1999.

[8] S.Tavazoie, J.D.Hughes, M.J.Campbell, R.J.Cho and G.M.Church, \Systematic determination of genetic network architecture", Nature Genetics, 22, 3, pp.281{285, 1999.