

Blind Gene Classification – An Application of a Signal Separation Method

Gen Hori¹

hori@bsp.brain.riken.go.jp

Masato Inoue^{2,3}

minoue@brain.riken.go.jp

Shin-ichi Nishimura^{2,4}

nishi@mns.brain.riken.go.jp

Hiroyuki Nakahara²

hiro@brain.riken.go.jp

- ¹ Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN
2-1, Hirosawa, Wako-shi, Saitama 351-0198, Japan
- ² Laboratory for Mathematical Neuroscience, Brain Science Institute, RIKEN
2-1, Hirosawa, Wako-shi, Saitama 351-0198, Japan
- ³ Department of Otolaryngology, Graduate School of Medicine, Kyoto University
54, Kawara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan
- ⁴ Department of Otolaryngology, Faculty of Medicine, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Keywords: gene classification, gene expression patterns, ICA(independent component analysis)

1 Introduction

Gene classification is one of important issues in gene expression data analysis because it is a basis for prediction of the functions of unknown genes. Cluster analysis is one of the recognized means of automatic gene classification. The present study shows that a new method based on ICA is a promising approach to automatic gene classification.

ICA(independent component analysis) is a recently developed method of signal processing and is mainly used for signal separation. It has been well established and successfully applied to analyze brain signals (such as EEG, MEG and MRI) and speech signals. Although ICA is similar to PCA(principal component analysis), ICA has some advantage to PCA because it exploits higher order statistics and has no restriction to orthogonal transformations.

The framework of ICA is as follows. Consider zero-mean and independent source signals $s(t) = (s_1(t), \dots, s_n(t))^t$ and assume that we observe the linear mixture of the source signals $x(t) = As(t)$. ICA attempts to find the de-mixing matrix W such that the de-mixed signals $y(t) = Wx(t)$ are as close to the source signals $s(t)$ as possible (excepting some indeterminacy) without using any knowledge on the mixing matrix A and the source signals $s(t)$. In our application of ICA to gene classification, $x(t)$ for each time point t corresponds to each gene profile from gene expression data.

2 Method and Results

For illustration of the validity of our new method, we used a gene expression data of yeast sporulation collected by Chu *et al.* [2] and available in public [4]. 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). To classify yeast genes based on the gene expression data, Chu *et al.* [2] hand-picked seven small subsets of representative genes using their domain knowledge on yeast genome and defined seven model profiles by averaging profiles of these genes in each subset (Fig.1). All the other yeast genes were classified into one of the seven groups whose model profile shows the highest correlation coefficient to the profile of each gene.

We applied ICA to the same gene expression data using the de-mixing model $y(t) = Wx(t)$ where $x(t)$ is a 7-dimensional vector which represents each gene profile. We used the JADE algorithm[1] to obtain the 7×7 de-mixing matrix W . Fig.2 gives the column plot of the inverse W^{-1} of the obtained de-mixing matrix. We denote the columns of W^{-1} by IC_i ($i = 1, 2, \dots, 7$) and regard them as model profiles. Notably, some model profiles in Fig.2 automatically generated by ICA are similar to those in Fig.1 obtained manually using domain knowledge on yeast genome.

To compare the ICA-generated model profiles with the manually obtained model profiles, we used only top three model profiles obtained by applying ICA to the pre-processed data reconstructed from the first three principal components. See [3] for the detail of the procedure. First, we sorted all the yeast genes according to their correlation coefficients and made a subset consists of the top 100 genes for each model profile. The left half of Table1 shows the numbers of genes in the intersections between these subsets. Second, we sorted all the genes according to the values of each separated component $y_i(t)$ corresponded to each profile and made subsets consist of the top 100 genes. The right half of Table1 shows the numbers in the intersections between these subsets.

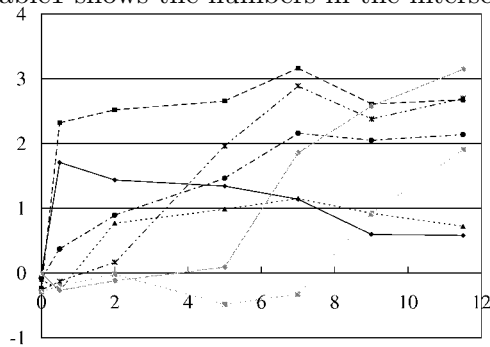


Fig.1 Model profiles obtained by Chu *et al.* [2].

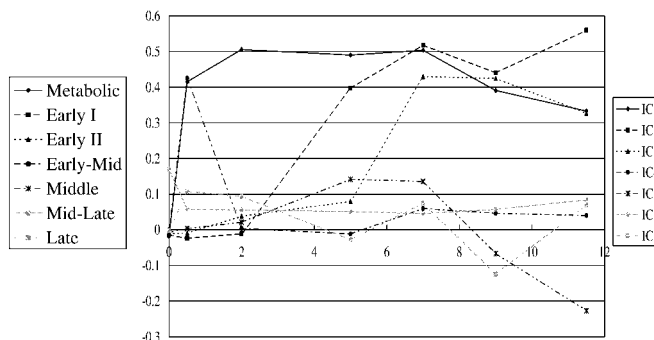


Fig.2 Column plot of W^{-1} .

3 Conclusion

Regarding the model profiles obtained by Chu *et al.*[2] as a benchmark, we conclude that the classified groups by our ICA-based method have a good match with the classified groups based on manually obtained model profiles. It is notable that our ICA-based method does not require a domain knowledge on genome and automatically classifies genes without any manual labor. Provided a rapid increase in use of microarray in experiments, this is a strong advantage.

	correlation			component		
	IC ₁	IC ₂	IC ₃	IC ₁	IC ₂	IC ₃
Metabolic	0	0	13	0	5	6
Early I	0	7	0	0	18	9
Early II	0	0	0	0	10	0
Early-Mid	19	0	0	10	1	0
Middle	54	0	0	49	1	0
Mid-Late	7	0	0	13	0	1
Late	0	0	0	1	0	4

Table1 Numbers of genes in the intersections

References

- [1] Cardoso, J.-F. and Souloumiac, A., Blind beamforming for non Gaussian signals, *IEEE Proc.-F*, 140:362–370, 1993.
- [2] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P., and Herskowitz, I., The transcriptional program of sporulation in budding yeast, *Science*, 282:699–705, 1998.
- [3] Hori, G., Inoue, M., Nishimura, S., and Nakahara, H., Blind gene classification based on ICA of microarray data, *Proc. ICA2001*, San Diego, USA, 2001.
- [4] <http://cmgm.stanford.edu/pbrown/sporulation>