Clustering Method Based on Onset and Cessation of Gene Expression

Kazumi Hakamada hakamada@brs.kyushu-u.ac.jp
Taizo Hanai taizo@brs.kyushu-u.ac.jp
Masahiro Okamoto okahon@brs.kyushu-u.ac.jp

Laboratory for Bioinformatics, Graduate School of Systems Life Sciences, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: clustering, gene expression data, time course data, onset, cessation.

1 Introduction

Gene clustering is one of the most important analyses for DNA microarray data. Most of the clustering methods require the whole expression time course profiles, however, most critical information is time period of onset of each gene. In gene network or circuit, gene switches for expression are turned on or off according to their demand. It should be effective to cluster the gene expression data based on expression profile considering both onset and cessation of each gene, however, there are few reports on such clustering method. In this study, we estimated timing of gene expression onset and its cessation by differential equation model [5]. Since most parameters about mRNA transcription rate, mRNA decay rate, onset time of gene and cessation are all still unknown, we optimized these parameters to realize experimental data followed by detail discussion focused on timing of gene expression onset and cessation. In this study, 45 genes about sporulation of *Saccharomyces cerevisiae* were classified by their timing of gene expression onset. The clustered genes and estimated expression period was evaluated by biological knowledge obtained by Chu, S. *et al.* [2].

2 Method and Results

2.1 Data Processing

In this study, time course data from DNA microarray of *Saccharomyces cerevisiae* [2] was used to analysis. The samples were harvested at time 0, 0.5, 2, 5, 7, 9 and 11.5 hours and applied to the microarray experiment. Forty five genes related to sporulation were extracted [6] and applied to the clustering.

2.2 Method

Differential equation model of gene expression represented by equation (1) was reported by Chen, T. *et al.* [1] and it was applied to *Dictyostelium* by Roman, S. *et al.* [5]. In this study, we applied the same differential equation model to *Saccharomyces cerevisiae* data. Since a lot of parameters (*i.e.* mRNA transcription rate, mRNA decay rate, onset time of gene and cessation) are all still unknown, these parameters were optimized so that they may realize experimental data. We estimated a set of t_1, t_2, γ_i and S_i which minimizes the value of the objective function represented by equation (2), where $E_i(t)$ represents experimental expression data about time t and $A_i(t)$ is estimated gene expression about time t. Among these estimated parameters, we especially focused on the parameters related to timing of gene expression onset and cessation (t_1 and t_2, respectively). We clustered these genes based on their onset and compared clustered genes with biological knowledge [2].

$$Ai(t) = \begin{cases} 0 & t \leq t_1^i \\ \dfrac{S_i}{\gamma_i}[1 - exp\{-\gamma_i(t - t_1^i)\}] & t_1^i < t \leq t_2^i \\ \dfrac{S_i}{\gamma_i}[1 - exp\{-\gamma_i(t_2^i - t_1^i)\}] & t_2^i < t \end{cases} \quad (1) \qquad error_i = \frac{\sum[E_i(t) - A_i(t)]^2}{\sum[E_i(t) - 1]^2} \quad (2)$$

where S_i: mRNA transcription rate, γ_i: mRNA decay rate, i: gene number, t_1: onset time, t_2: cessation time.

2.3 Result

Figure 1 shows the calculated and experimental time course data. Calculated time course data shows little discrepancy with experimental data ($error = 3\times10^{-4}$), however, the average value for *error* about 45 genes is very low (4.1×10^{-3}). Based on gene expression onset, 45 genes were divided into 4 clusters. As the result, cluster 1(C1), cluster 2(C2), cluster 3(C3) and cluster 4(C4) had 25, 3, 14 and 3 genes, respectively (Fig. 2). Chu, S. *et al.* had reported that 21 genes among these 45 genes could be experimentally classified into the following temporal expression class: Early1, Early2, Early-Middle, Middle, Middle-Late [2]. Table 1 shows the percentage of agreement between our classification and classification by Chu, S. *et al.* Our category C1 includes all Early1 category by Chu, S. *et al.*, and most of all Early2 genes (87.5%). C2 covers 66.7% Early-Middle genes and C3 does most of all Middle genes (91.7%). C4 does 66.7% of Middle-Late genes. Thus the order of expression onset in our categories, C1∼C4, was completely consistent with that of categories by Chu, S. *et al.* [2].

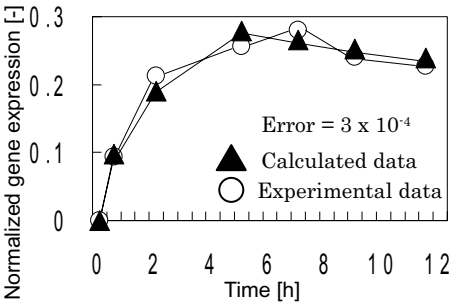

Figure 1: Inferred data and experimental data.

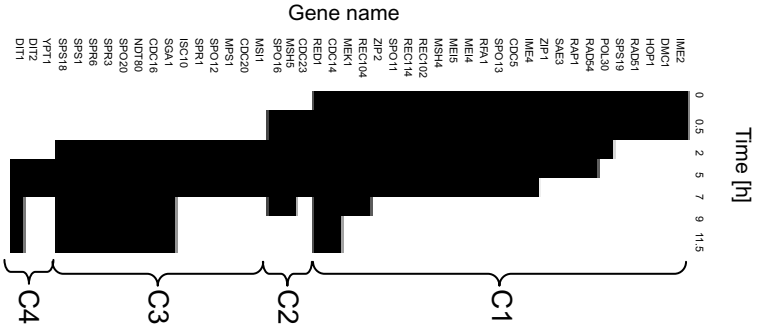

Figure 2: Gene expression onset and cessation.

Table 1:

	Early1	Early2	Early-Middle	Middle	Middle-Late	Others
C1	100.0	87.5	33.3	8.3	0.0	62.5
C2	0.0	12.5	66.7	0.0	0.0	0.0
C3	0.0	0.0	0.0	91.7	33.3	25.0
C4	0.0	0.0	0.0	0.0	66.7	12.5

3 Discussion

For the validation of our method, another kind of data related cell cycle [4] were analyzed in the same manner. Estimated expression time course and the timing for gene expression onset and cessation of most genes were also consistent with biological knowledge obtained by Etienne, S. *et al.* [3]. These results show the wide applicability of our method as a general clustering.

In this study, we have made clustering of genes focusing only onset time of each gene. The other clustering based on cessation time, S_i, γ_i, could be applicable as a future study.

References

[1] Chen, T., He, H.L., and Church, G.M., Modeling gene expression with differential equations, *Proc. Pacific Symp. Biocomputing '99, World Scientific*, 29–40, 1999.

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

[3] Etienne, S. and Kim, N., *CLB5* and *CLB6*, a new pair of B cyclins involved in DNA replication in *Saccharomyces cerevisiae*, *Genes & Dev.*, 7:1160–1175, 1993.

[4] Raymond, J.C., Michael, J.C., Elizabeth, A.W., Lars, S., Andrew, C., Lisa, W., Tyra, G.W., Andrei, E.G., David, L., David, L., and Ronald, W.D., A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell*, 2:65–73, 1998.

[5] Roman, S., Negin, I., Terence, H., and William, F.L., Extracting transcriptional events from temporal gene expression patterns during *Dictyostelium* development, *Bioinformatics*, 18:61–66, 2002.

[6] Tomida, S., Hanai, T., Honda, H., and Kobayashi, T., Analysis of gene expression profile using fuzzy adaptive resonance theory, *Bioinformatics*, 18:1073–1083, 2002.