

# ASIAN: Automatic System for Inferring a Network from Gene Expression Profiles

Katsuhisa Horimoto<sup>1</sup>

horimoto@post.saga-med.ac.jp

Hiroyuki Toh<sup>2</sup>

toh@beri.or.jp

<sup>1</sup> Laboratory of Mathematics, Saga Medical School, Nabeshima 5-1-1, Saga 849-8501, Japan

<sup>2</sup> Department of Bioinformatics, Biomolecular Engineering Research Institute, Furuedai 6-2-3, Suita, Osaka 565-0874, Japan

**Keywords:** hierarchical clustering, genetic network, graphical Gaussian modeling

## 1 Introduction

Recently, we have developed two methods for expression profile analysis. One is the automatic determination of cluster boundaries in hierarchical clustering of profiles [2], and another is the inference of a genetic network by application of graphical Gaussian modeling (GGM) [4]. Here, we synthesize the newly developed methods into a system, named **ASIAN** (**A**utomatic **S**ystem for **I**nferring **A** Network). The performance is illustrated by means of yeast profile data. .

## 2 Method and Results

### 2.1 Profile Data

The gene expression profile data analyzed here are cited from Spellman et al. (1998) [3]. From the expression profiles of 6178 yeast (*Saccharomyces cerevisiae*) genes that were measured under 77 conditions, 2918 profiles with no or fewer than one missing data in 77 conditions are selected.

### 2.2 Algorithm

**Step 1:** A metric between profiles is defined as the Euclidean distance between correlation coefficients, i.e.,  $d_{ij} = \sqrt{\frac{1}{n} \sum_{k=1}^n (p_{ik} - p_{jk})^2}$ , where  $p_{ik}$  is Pearson's correlation coefficient between  $i$  and  $k$  gene profiles and  $n$  is the total number of genes.

**Step 2:** A standard hierarchical clustering technique, the group average method, is applied to the distance matrix calculated by **Step 1**.

**Step 3:** According to the dendrogram at the nodes from 2918 to 2, the variation inflation factor (VIF), which is expressed as  $VIF = r_{ii}^{-1}$ , where  $r_{ii}^{-1}$  is  $i$ th diagonal element of the inverse of correlation coefficient matrix, is calculated from the matrix that is composed of representative profiles. The calculation of VIF is started at the nodes in ascending order of the dendrogram, and is stopped, when all of VIF's are less than 10.0 that is a familiar cut-off value [1].

**Step 4:** For the clusters determined by **Step 3**, the expression levels at each condition are averaged over the members of the cluster, and the correlation coefficient between every pair of clusters is calculated for the average profiles.

**Step 5:** The correlation coefficient matrix obtained in **Step 4** is analyzed by GGM [5]. When the probability of the deviance between the matrices at one step and the preceding step in the iterative analysis is less than 0.05, the iteration is stopped, and a partial correlation coefficient matrix for the clusters is obtained.

### 2.3 Number of Clusters Determined by *ASIAN*

As seen in Fig. 1, the fraction of VIF's more than 10.0 monotonously decreases from 70 clusters, and finally reached at 0 value at 30 clusters. The monotonous decrease indicates no possibility that the fraction has 0 value in more than 70 clusters. In contrast, the fraction holds 0 value in less than 30 clusters, suggesting that the clusters are clearly separated in these clusters. Thus, 2918 profiles were automatically grouped into 30 clusters. The number of members of each cluster was ranged from 5 to 308.

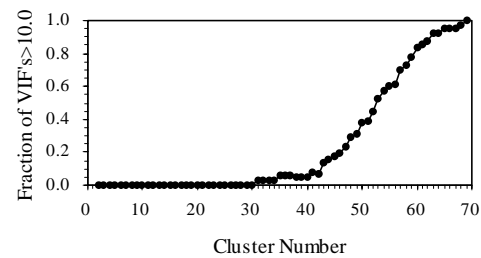


Figure 1: Determination of cluster boundaries. The fraction of VIF's more than 10.0 is plotted against the cluster number.

### 2.4 Network Inferred by *ASIAN*

The network inferred by *ASIAN* is schematically shown in Fig. 2. In 435 connections between 30 clusters, 161 connections (37.0%) were broken off by GGM. Thus, 274 connections between 30 clusters (63.0%) were established by *ASIAN*. The maximum number of connections is 23 in cluster 25, and the minimum number of connections is 13 in cluster 20. Notably, the numbers of members belonging to the two clusters are 175 and 128, respectively. The number of members, therefore, appears to be independent of whether a cluster is connected with or not. Rather, it is considered that the establishment of the connection depends on the relationship between the genes, which is expressed by the profiles.

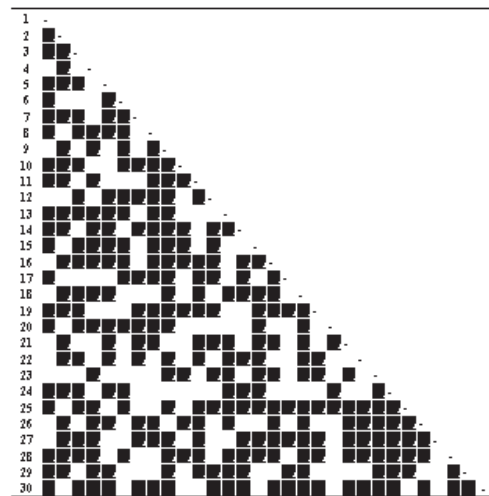


Figure 2: Inferred Network between 30 Clusters. The open circle indicates no connection between the two clusters, and the closed cell indicates the connection.

## 3 Discussions

The members of 30 clusters were classified in view from a classification scheme of gene function. Furthermore, the present network will be corresponded with the known regulatory networks.

## Acknowledgement

K. H. was supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Science, Sports, and Culture of Japan (grant 13208028).

## References

- [1] Freund, R.J. and Wilson, W.J., *Regression Analysis*, Academic Press, 1998.
- [2] Horimoto, K. and Toh, H., Statistical estimation of cluster boundaries in gene expression profile data, *Bioinformatics*, in press.
- [3] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell.*, 9:3273–3297, 1998.
- [4] Toh, H. and Horimoto, K., Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, in press.
- [5] Whittaker, J., *Graphical Models in Applied Multivariate Statistics*, John Wiley, 1990.