

Modification of ASIAN: Deduction of a Framework of Gene Regulatory Systems from Expression Profile Data

Katsuhisa Horimoto¹

khorimot@ims.u-tokyo.ac.jp

Sachiyo Aburatani²

sachiyo@grt.kyushu-u.ac.jp

Satoru Kuhara²

kuhara@grt.kyushu-u.ac.jp

Hiroyuki Toh³

toh@kuicr.kyoto-u.ac.jp

¹ Laboratory of Biostatistics, Human Genome Center, Institute of Medical Science,

University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

² Laboratory of Molecular Gene Technics, Graduate School of Genetic Resources Technology, Kyushu University, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan

³ Laboratory of Genome Informatics, Bioinformatics Center, Institute of Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: gene classification, graphical Gaussian modeling, network comparison

1 Introduction

Recently, we developed an automatic system for deducing a framework of regulatory relationships from gene expression data, named ASIAN [1, 2]. One of the merits of our system is that it simultaneously performs the gene classification and the relation inference from a large amount of gene expression data by combining the graphical Gaussian modeling with standard multivariate statistical techniques.

In this work, we describe our modifications of the previous ASIAN to estimate the cluster boundaries, by setting a user-defined threshold for measuring the linear relationship between the profiles of clusters. The feasibility of the modified ASIAN is demonstrated by the comparison between two frameworks inferred from two sets of clusters that were estimated by distinctive thresholds.

2 Method and Results

2.1 Expression Profile Data

The gene expression profile data analyzed here were cited from Gasch *et al.* [3]. The data comprise the expression profiles of 6152 yeast (*Saccharomyces cerevisiae*) genes that were measured under 173 conditions.

2.2 Procedure of ASIAN

A basic idea of automatic gene classification is estimation of linear relationship between gene expression profile vectors whose elements are composed of mRNA expression levels in the monitored conditions. To estimate the linear relationship, the variance inflation factor (VIF) in the multi-regression analysis is adopted. The profile vectors are estimated to be independent if all VIFs' are less than a threshold value. By the calculation of VIF in ascending order of nodes along the dendrogram, we will find the maximum number of clusters in that the above condition is satisfied. Then the graphical Gaussian modeling (GGM) is applied to infer a network between the gene clusters.

2.3 Modification of ASIAN

In a new version of ASIAN, we designed a user-defined threshold for VIF in estimating the cluster number. The new modification of the threshold for VIF is expected to increase the performance of the system. First, the variety of cluster numbers enables us to compare several sets of connections between clusters. The networks between the clusters can be investigated from a hierarchical viewpoint with any amount of biological knowledge on the gene classes and the gene regulatory systems. Secondly, since the profile data sets are composed of different numbers of genes, measured under different conditions from different organisms, the user-defined threshold for VIF is useful for explanatory analyses of the profile data. Thirdly, since the cluster number increases as the VIF is set to a larger value, a variable setting of the VIF threshold also serves to estimate the maximum number of clusters derived from the profile data analyzed. Although one parameter for the threshold of VIF is added to the system, the improvement enables us to perform a heuristic search for the gene classification and the gene regulatory systems with reference to the biological information.

2.4 Application to a Whole Set of Yeast Gene Expression Profiles

The cluster numbers estimated with six thresholds for VIF by the modified ASIAN are plotted in Fig. 1. The cluster numbers increase as the thresholds for VIF increase. In particular, the curve of the thresholds for VIF versus the cluster number saturates in the threshold of about 50.0: the cluster number is 60 when the threshold for VIF is 50.0 and the number is 61 when the threshold is 100.0. The saturation of the curve suggests a limitation of the division into the clusters that mutually show no linear relationship. With these situations in mind, the genes are classified into 30 and 60 clusters in the present study.

In the 30 clusters, the numbers of members range from 6 to 727, while they are from 3 to 727 in the 60 clusters. The detailed comparison between the two sets of clusters will be discussed in terms of the gene function in the poster.

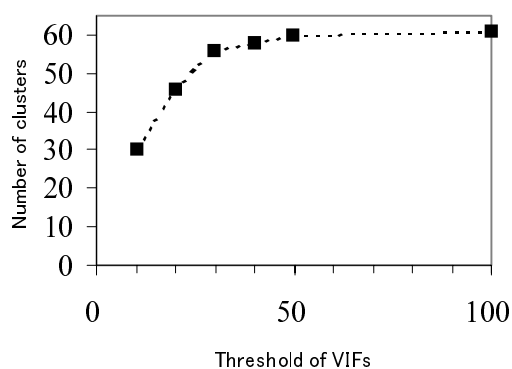


Figure 1: Cluster numbers estimated by user-defined thresholds for VIF.

3 Discussion

The improvement of the gene classification, by setting a user-defined threshold for VIF, allows the user to compare frameworks between different clusters, which are inferred by GGM. The detailed comparison between the two inferred networks will be discussed in the poster, in terms of the known regulatory system.

References

- [1] Aburatani, S., Kuhara, S., Toh, H., and Horimoto, K., Deduction of a gene regulatory relationship framework from gene expression data by the application of graphical Gaussian modeling, *Signal Processing*, in press.
- [2] Horimoto, K., Aburatani, S., Kuhara, S., and Toh, H., ASIAN - automatic system for inferring a network from gene expression profiles, *Res. Commun. Biochem. Cell Mol. Biol.*, in press.
- [3] <http://genome-www5.stanford.edu/MicroArray/SMD/>