Finding Sparse Gene Networks

Ayumi Shinohara[1] ayumi@i.kyushu-u.ac.jp
Keisuke Iida[1] k-iida@i.kyushu-u.ac.jp
Masayuki Takeda[1] takeda@i.kyushu-u.ac.jp
Osamu Maruyama[2] om@math.kyushu-u.ac.jp
Satoru Miyano[3] miyano@ims.u-tokyo.ac.jp
Satoru Kuhara[4] kuhara@grt.kyushu-u.ac.jp

[1] Department of Informatics, Kyushu University 33, Fukuoka 812-8581, Japan
[2] Graduate School of Mathematics, Kyushu University, Fukuoka 812-8581, Japan
[3] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
[4] Graduate School of Genetic Resources Technology, Kyushu University, Fukuoka 812-8581, Japan

Keywords: gene network, graphical model, DNA microarray

1 Introduction

DNA microarray technology enabled us to produce time series of gene expression patterns. Our research group launched a project whose purpose is to reveal the gene regulatory networks among the 6,200 genes of *Saccharomyces cerevisiae*.

We have introduced a *weighted network model* as an edge-weighted graph, where each weight reflects the strength of the interaction, and analyzed its computational complexity [4]. We also proposed an algorithm to adjust the weights incrementally. Based on the algorithm, we have been developing a system to find genetic networks and visualize them [3].

One of the most serious problem in our model is that the network produced by our system might be *dense*, since we have not put any restrictions on in-degree nor out-degree of the network, although many biologists claim that the network should be *sparse*. We need some methods to reduce the edges so that the resulting network is sparse enough.

In this paper, we propose a method to reduce the edges inspired by the covariance selection models [2].

2 Method and Results

Basically, we follow the iterative algorithm that infers important covariances in covariance selection model [2].

Let $R = (r_{ij})$ be the *correlation coefficient matrix* that is calculated by the gene expression profile data. From R, we compute the *partial correlation coefficient matrix* $P = (p_{ij})$ by $p_{ij} = -\frac{r^{ij}}{\sqrt{r^{ii}}\sqrt{r^{jj}}}$, where $R^{-1} = (r^{ij})$ is the inverse matrix of R. Fig. 1 shows a *hybrid matrix* of both correlation and partial correlation coefficient matrices, where the correlation coefficients are shown in the lower left triangle, and the partial correlation coefficients in the upper right triangle. Then we find a non-zero partial correlation coefficient whose absolute value is the minimum among them, and force it to be 0. The corresponding correlation coefficient can be recalculated efficiently (see [2]). Repeat this procedure until the *deviance* of the matrix exceeds the given threshold value. In this way, we can reduce the edges based on the statistical theory.

However, in practice, we cannot treat a large matrix (more than, say, 35), because of the *multicollinearity*, due to the existence of high correlations among the variables. Since the determinant of

$$\begin{pmatrix} 1 & -0.18 & 0.43 & 0.12 & -0.32 \\ -0.22 & 1 & 0.08 & -0.19 & -0.10 \\ 0.47 & -0.12 & 1 & 0.36 & 0.32 \\ 0.35 & -0.25 & 0.46 & 1 & -0.04 \\ -0.18 & 0.05 & 0.21 & 0.03 & 1 \end{pmatrix} \Longrightarrow \begin{pmatrix} 1 & -0.18 & 0.43 & 0.12 & -0.32 \\ -0.22 & 1 & 0.08 & -0.19 & -0.10 \\ 0.47 & -0.12 & 1 & 0.36 & 0.32 \\ 0.35 & -0.25 & 0.46 & 1 & \mathbf{0} \\ -0.18 & 0.05 & 0.21 & \mathbf{0.06} & 1 \end{pmatrix}$$

deviance 0.0 deviance 0.080

Figure 1: An example of hybrid matrix. The right matrix can be get by forcing $p_{45} = -0.04$ to be 0, at the cost of deviance becomes 0.080 from 0.0.

the matrix becomes too small, round-off errors and overflow errors cause some troubles. We overcome this difficulty as follows. We diagnose the matrix as multicollinear when a *variance inflation factor (VIF)* [1] is larger than a cutoff value, usually 10.0. We define the i-th VIF of a correlation coefficient matrix R by $VIF_i = r^{ii}$, where r^{ii} is the i-th diagonal element of the inverse matrix of R. VIF expresses the degree of linear relationship between the profile data [1].

Unfortunately, in our gene expression profile data, we have observed that the VIF exceeds 10.0 for any combinations of more than 35 genes. It implies that we can treat at most 35 genes simultaneously. We use the above method for the following two applications.

1. Construct a large network by partial constructions: Pick up 35 genes randomly and get a sparse network by the above method. Repeat this procedure enough and construct a whole network by merging these partial networks.

2. For a set S of genes ($|S| < 35$) that are specified by a user, construct a partial network for S: Pick up additional genes T randomly such that $|S \cup T| \leq 35$, and get a sparse network by the above method. Repeat this procedure enough and construct a network for S by "averaging" these networks.

We are currently executing the experiments, and we will show the results at the poster.

Acknowledgements

The authors thank Prof. K. Horimoto for his helpful advice on covariance selection. This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Area 'Genome Science' from the Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Horimoto, K. and Toh, H., A Procedure for estimating cluster boundaries in gene expression profile data, *Genome Informatics*, 11, 2000.

[2] Miyakawa, M., *Graphical Modeling* (in Japanese), Asakura-Shoten, 1997.

[3] Moriyama, T., Shinohara, A., Takeda, M., Maruyama, O., Goto, T., Miyano, S., and Kuhara, S., A system to find genetic networks using weighted network model, *Genome Informatics*, 10:186–195, 1999.

[4] Noda, K., Shinohara, A., Takeda, M., Matsumoto, S., Miyano, S., and Kuhara, S., Finding genetic network from experiments by weighted network model, *Genome Informatics*, 9:141–150, 1998.