Multiscale Bootstrap Analysis of Gene Networks Based on Graphical Gaussian Modeling

Takeshi Kamimura[1] **Hidetoshi Shimodaira**[1]

kamimur1@is.titech.ac.jp shimo@is.titech.ac.jp

[1] Department of Mathematical and Computing Sciences, Tokyo Institute of Technology,
Ookayama, Meguro, Tokyo 152-8552, Japan

Keywords: multiscale bootstrap, AU test, graphical Gaussian model, microarray, gene network

1 Introduction

The development of DNA microarray technology has enabled gene expression analysis based on simultaneous observation of thousands of genes. One of the purposes of microarray analysis is to estimate the relationships among genes, and thus exploratory method such as cluster analysis or graphical modeling is often used. However, the estimation is often susceptible by statistical sampling error, and thus the result is obtained only by chance without reflecting the true hypothesis. Therefore, it is necessary to evaluate the reliability of hypothesis obtained as the result of analysis.

In this study, we present the program, which is written in R language, to assess the confidence of the gene network based on graphical Gaussian model by giving the p-value for the edges connecting genes. The main thrust of the program is to calculate the p-value of the Approximately Unbiased (AU) test using the multiscale bootstrap resampling [4, 5, 6]. This method was developed recently to improve the accuracy of the bootstrap probability, and has been used widely in phylogenetic analysis.

2 Methods

2.1 Graphical Gaussian Model (GGM)

Graphical Gaussian model, known as covariance selection model, assumes multivariate normal distribution for observed data and measures conditional independence relationships between two random variables based on partial correlation coefficient. If the two variables are conditionally independent, they are not connected by an edge, otherwise they are connected. In this way an undirected graph is constructed to represent dependence among variables. By applying this method to microarray expression data, we can obtain a gene network which shows dependence among genes as an undirected graph. The details of estimation of gene network based on graphical Gaussian model are described in [7].

2.2 Bootstrap and Multiscale Bootstrap Edge Intensity

We measure the intensity of the edge by the bootstrap and multiscale bootstrap method. In the multiscale bootstrap method, we generate replicates $\boldsymbol{X}^*_{n'} = (\boldsymbol{x}^*_1, \dots, \boldsymbol{x}^*_{n'})$ for several n' values from the original gene expression data $\boldsymbol{X}_n = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. In other words, we alter the number of arrays from n to n' in the bootstrap replication. We will take n' values with $n'/n = 0.5,0.6,0.7,0.8,0.9,1.0,1.1,1.2,1.3,1.4$, in the example shown later. We call $\tau = \sqrt{n/n'}$ scale. The bootstrap algorithm with n' arrays can be expressed as follows:

Step1: Generate the bootstrap replicate $\boldsymbol{X}^*_{n'}$.
Step2: Estimate the gene network from $\boldsymbol{X}^*_{n'}$.
Step3: Iterate Step1 and Step2 B times. Then we obtain B gene networks.
Step4: If the edges $gene_i \leftrightarrow gene_j$ exist $k(\tau)$ times in the B networks, we then define the bootstrap edge intensity between $gene_i$ and $gene_j$, $BP_{ij}(\tau)$, as $k(\tau)/B$.

In the ordinary boostrap method, we take $n' = n$, and the bootstrap edge intensity can be written as $BP_{ij}(1)$. In the multiscale bootstrap method, we calculate $BP_{ij}(\tau)$ with several τ values by altering

n'/n. Then we calculate the multiscale bootstrap edge intensity between $gene_i$ and $gene_j$ from $BP_{ij}(\tau)$ values. According to the statistical geometric theory of Efron *et al.* [1] and Shimodaira [4], the very accurate probability value is expressed as $AU_{ij} = 1 - \Phi(d_{ij} - c_{ij})$ using geometric quantities d_{ij} and c_{ij}, where Φ denote the ditribution function of standard normal distribution. We estimate d_{ij} and c_{ij} by fitting the theorerical curve $BP_{ij}(\tau) = 1 - \Phi(d_{ij}/\tau + c_{ij}\tau)$ to the observed $BP_{ij}(\tau)$ values calculated by the multiscale bootstrap method.

3 A Numerical Example

In our program, once you select the microarray data to analyze and give the number of bootstrap replicates, all the multiscale bootstrap edge intensities in the network are calculated automatically. Here we applied the program to the *S. cerevisiae* gene expression data. We focused on 9 genes, which are involved or putatively involved in the heat shock response. We took $B = 10,000$ for each scale. Table 1 shows the matrix of all the multiscale bootstrap edge intensities and only edges with high multiscale bootstrap intensities are shown in Figure 1. The estimated network captured the dependence of HSF1 and SSA1, SSA1 and SSA3, which is reported in [2, 3].

Table 1: Matrix of the multiscale bootstrap edge intensities. $(AU_{ij},\ 1 \le j < i \le 9)$

HSP104	–								
HSF1	0.842	–							
HIG1	0.982	0.969	–						
MCM1	0.884	0.632	0.999	–					
SSA1	0.885	0.951	0.978	0.275	–				
SSA3	0.469	0.595	0.714	0.996	1.000	–			
HSP26	0.993	0.978	0.991	0.601	0.399	0.305	–		
HSP82	0.999	0.854	0.323	0.543	1.000	0.353	0.979	–	
YRO2	0.542	0.943	0.514	0.926	0.972	0.956	0.609	0.408	–
	HSP104	HSF1	HIG1	MCM1	SSA1	SSA3	HSP26	HSP82	YRO2

We developed a program for the analysis as an add-on package for a statistical package R. It will be available at our website [8].

This work is supported in part by Grant-in-Aid for Young Scientists (A) KAKENHI-14702061 from MEXT of Japan.

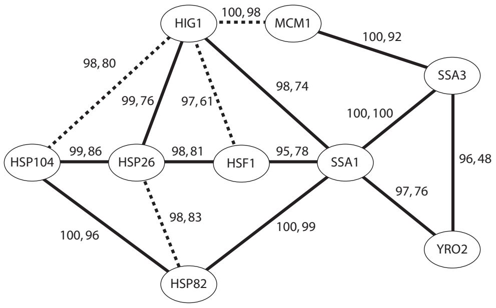

Figure 1: Estimated gene network. Only edges with AU_{ij} larger than 0.95 are shown. The numbers next to the line are AU_{ij} (left) and BP_{ij} (right) in percentage. Note that dashed lines indicate a negative partial correlation and solid lines indicate a positive partial correlation.

References

[1] Efron, B., Halloran, E., and Holmes, S., Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA,* 93:13429–13434, 1996

[2] Glover, J.R. and Lindquist, S., Hsp104, Hsp70, and Hsp40: a novel chaperone system that rescues previously aggregated proteins. *Cell*, 94:73–82, 1998.

[3] Shi, Y., *et al.*, Molecular chaperones as HSF1-specific transcriptional repressors. *Genes* & *Development*, 12:654–666, 1998.

[4] Shimodaira, H., An approximately unbiased test of phylogenetic tree selection, *Systematic Biology*, 51:492–508, 2002.

[5] Shimodaira, H. and Hasegawa, M., CONSEL: for assessing the confidence of phylogenetic tree selection, *Bioinformatics*, 17:1246–1247, 2001.

[6] Shimodaira, H. (in press), Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, *Annals of Statistics*, 2004.

[7] Toh, H. and Horimoto, K., Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, 18:287–297, 2002.

[8] http://www.is.titech.ac.jp/~shimo/prog/