Multiscale Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression

Takeshi Kamimura[1]
kamimur1@is.titech.ac.jp

Hidetoshi Shimodaira[1]
shimo@is.titech.ac.jp

Seiya Imoto[2]
imoto@ims.u-tokyo.ac.jp

SunYong Kim[2]
sunk@ims.u-tokyo.ac.jp

Kousuke Tashiro[3]
ktashiro@grt.kyushu-u.ac.jp

Satoru Kuhara[3]
kuhara@grt.kyushu-u.ac.jp

Satoru Miyano[2]
miyano@ims.u-tokyo.ac.jp

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

[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[3] Guraduate School of Genetic Resources Technology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: multiscale bootstrap, Bayesian network, nonparametric regression, microarray

1 Introduction

The Bayesian network [2, 3, 4] is a very powerful tool for estimating the gene network from microarray expression profiles. The estimated network is often susceptible to statistical sampling error, and thus Imoto *et al.* [3, 4] evaluated the reliability of estimation by calculating the bootstrap probabilities for the edges connecting genes. The bootstrap method, however, underestimates the probability values, and it sometimes leads to false "discovery". For improving the accuracy of the bootstrap probability, we propose the application of the newly developed multiscale bootstrap [5, 6] to the gene network estimation.

2 Method

2.1 Nonlinear Bayesian Network Model

In the estimation of a gene network, Imoto *et al.* [3, 4] proposed the nonlinear Bayesian network model for caputuring even nonlinear relationship among genes by using the nonparametric regression model. The criterion, BNRC, was newly introduced for evaluating the estimated gene network from Bayes approach. The details of the nonlinear Bayesian network model are described in [4].

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 $\boldsymbol{X}_{n'}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{n'}^*)$ for several n' values from the original gene expression data $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. 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 $gene_i \to gene_j$ and $gene_j \to gene_i$ exist $k_1(\tau)$ and $k_2(\tau)$ times, respectively, in the B networks, we then define the bootstrap edge intensity between $gene_i$ and $gene_j$, $BP_{ij}(\tau)$, as $(k_1(\tau) + k_2(\tau))/B$.

In the normal 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 [5], the very accurate probability value is expressed as $MS_{ij} = 1 - \Phi(d_{ij} - c_{ij})$ 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 Result

We applied the proposed method 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 $B = 10000$. Figure 1 is the resulting network. The multiscale bootstrap edge intensity MS_{ij} is shown by the line width, and the number next to the line is the degree of confidence of the edge direction. The observed $BP_{ij}(\tau)$ values and the fitted curves are shown for some of the edges in Figure 2. Table 1 shows the gene pairs with high multiscale bootstrap intensities.
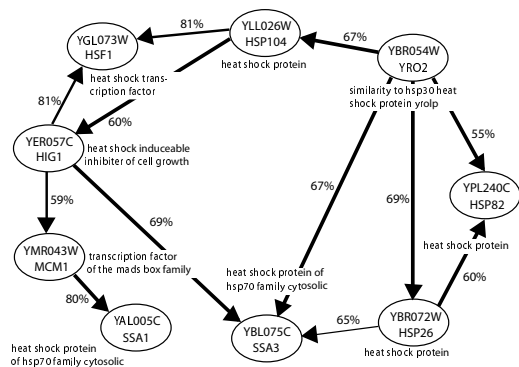
Table 1: Gene pairs with high multiscale bootstrap intensities.

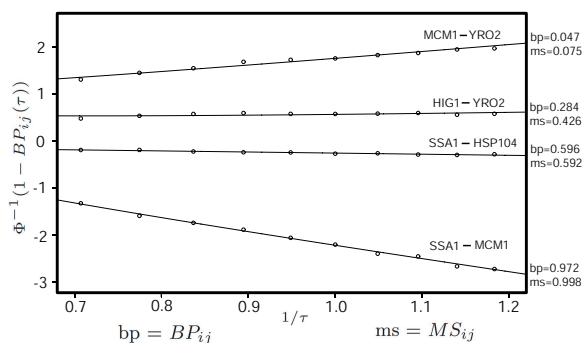| $gene_i$	$gene_j$	MS_{ij}	BP_{ij}
YBR072W	YBR054W	1.000	0.985
YBR072W	YPL240C	1.000	0.998
YER057C	YLL026W	0.998	0.999
YAL005C	YMR043W	0.998	0.972
YLL026W	YBR054W	0.996	0.938
YER057C	YBL075C	0.994	0.985
YPL240C	YBR054W	0.991	0.835
YBL075C	YBR054W	0.978	0.885
YER057C	YGL073W	0.928	0.642
YGL073W	YLL026W	0.918	0.922
YER057C	YMR043W	0.915	0.851

Figure 1: The resulting network.

Figure 2: The curve fitting to the observed $BP_{ij}(\tau)$ values.

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] Friedman, N., Linial, M., Machman, I., and Pe'er, D., Using Bayesian networks to analyze expression data, *J. Comp. Biol.*, 7:601–620, 2000.

[3] Imoto, S., Goto, T., and Miyano, S., Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression, *Pac. Symp. Biocomput., World Scientific*, 7:175–186, 2002.

[4] Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., and Miyano, S., Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *J. Bioinformatics and Computational Biology*, 1(2):231–252, 2003.

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

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