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

Seiya Imoto[1] imoto@ims.u-tokyo.ac.jp
SunYong Kim[1] sunk@ims.u-tokyo.ac.jp
Hidetoshi Shimodaira[2] shimo@is.titech.ac.jp
Sachiyo Aburatani[3] sachiyo@grt.kyushu-u.ac.jp
Kousuke Tashiro[3] ktashiro@ims.u-tokyo.ac.jp
Satoru Kuhara[3] kuhara@ims.u-tokyo.ac.jp
Satoru Miyano[1] miyano@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[2] Department of Mathematical and Computing Sciences, Tokyo Institute of Technology, Ookayama, Meguro, Tokyo 152-8552, Japan
[3] Graduate School of Genetic Resources Technology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: Bootstrap, Bayesian network, nonparametric regression, microarray

1 Introduction

The development of the microarray technology provides us a huge amount of gene expression profiles. The estimation of a gene network has received considerable attention in the field of bioinformatics and several methodologies have been proposed such as the Boolean network [1], the Bayesian network [3, 4, 5] and so on. In this paper, we propose the method for measuring the reliability of the estimated gene network by using the bootstrap method [2].

2 Method

2.1 Nonlinear Bayesian Network Model

In the estimation of a gene network, Imoto *et al.* [4, 5] proposed the nonlinear Bayesian network model for capturing 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 [5].

2.2 Bootstrap Edge Intensity and Degree of Confidence of Bayes Causality

We measure the intensity of the edge and the degree of confidence of the direction of the Bayes causality by the bootstrap method. The algorithm can be expressed as follows:
Step1: Make the bootstrap gene expression matrix $\boldsymbol{X}_n^* = (\boldsymbol{x}_1^*, ..., \boldsymbol{x}_n^*)^T$ by randomly sampling n times, with replacement, from the original gene expression data $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ of n microarrays.
Step2: Estimate the gene network from $\boldsymbol{X}_n^*$.
Step3: Repeat Step1 and Step2 T times.

From this algorithm, we obtain T gene networks. We define the bootstrap intensity of edge and direction of Bayes causality as follows: *Edge intensity*: If the edges $gene_i \rightarrow gene_j$ and $gene_j \rightarrow gene_i$ exist t_1 and t_2 times in the T networks, respectively, we then define the bootstrap edge intensity between $gene_i$ and $gene_j$ as $(t_1 + t_2)/T$. *Degree of confidence of the Bayes causality*: If $t_1 > t_2$, we adopt the direction $gene_i \rightarrow gene_j$ and define that the degree of confidence of causality is $t_1/(t_1 + t_2)$.

We superpose the bootstrap networks and original network. The superposed network contains edges which have small intensities. Therefore, we can set a certain threshold value and remove the edges whose intensities are under the threshold. We note that the superposed network possibly does not hold the acyclic assumption, but much effective information are in this network.

3 Result

We applied the proposed method to the *S. cerevisiae* gene expression data. We focused on 521 genes and used 100 gene disruption microarrays. The bootstrap algorithm was repeated 100 times. Figure 1 is the resulting partial network. The edge intensity is shown by the line width, and the number next to the line is the degree of confidence of the direction. Table 1 shows the gene pairs with high bootstrap intensities.

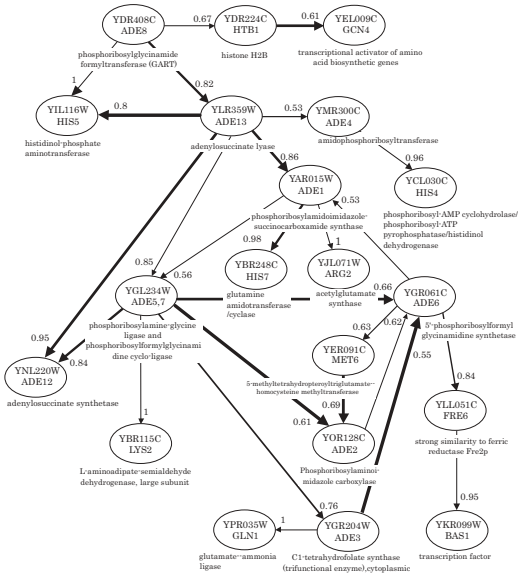

Figure 1: The resulting partial network.

Parent	*Child*	*Inte.*	*Dire.*	*Biological knowledge*
CUP1A	CUP1B	1.00	0.86	Related Proteins(100%)
GLK1	TPS1	1.00	0.78	No data
HHF1	HHF2	1.00	0.73	Related Proteins(100%)
HSC82	HSP82	1.00	0.61	Related Proteins(97%)
PHO11	PHO12	1.00	0.57	Related Proteins(100%)
ARO10	ARO9	1.00	0.57	Both ARO9 and ARO10 are transcriptionally regulated by Aro80p (34415)
ASP3A	ASP3C	1.00	0.52	Related Proteins(100%)
PHO5	PHO3	0.99	0.98	Related Proteins(87%)
HSP104	PMC1	0.99	0.91	Related Proteins(93%)
GAL11	SSN6	0.99	0.85	GAL11: polyglutamine and poly-glutamine-alanine domain are similar to those found in Ssn6p
FBA1	GPM1	0.99	0.83	Functional genomics
YOL002C	OLE1	0.99	0.59	Functional genomics
ADE3	ADE6	0.99	0.56	Functional genomics
IDH1	IDH2	0.99	0.54	Related Proteins(42%) : Protein-protein interaction
HAP1	TRK2	0.98	0.56	No data
HHF2	HTB1	0.97	0.89	Both relates to Histone
YDR516C	PPR1	0.97	0.78	No data
DBF4	CRZ1	0.97	0.60	No data
YNL134C	GRE2	0.97	0.56	Functional genomics
SME1	REG2	0.97	0.55	No data
PDR5	PDR15	0.97	0.55	Related Proteins(75%)
TPS2	HSP78	0.97	0.52	Functional genomics

Table 1: Gene pairs with high bootstrap intensities.

References

[1] Akutsu, T., Kuhara, S., Maruyama, O., and Miyano, S., Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions, *Proc. 9th ACM-SIAM Symp. Discrete Algorithms*, 695–702, 1998.

[2] Efron, B., Bootstrap methods: Another look at the jacknife, *Ann. Statist.*, 7:1–26, 1979.

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

[4] Imoto, S., Goto, T., and Miyano S., Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression, *Proc. Pacific Symposium on Biocomputing*, *World Scientific*, 7:175–186, 2002.

[5] 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, *Proc. IEEE Computer Society Bioinformatics Conference*, *Computer Society Press*, 219–227, 2002.