Dynamic Bayesian Network and Nonparametric Regression Model for Inferring Gene Networks

SunYong Kim sunk@ims.u-tokyo.ac.jp
Seiya Imoto imoto@ims.u-tokyo.ac.jp
Satoru Miyano miyano@ims.u-tokyo.ac.jp

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

Keywords: dynamic Bayesian network, nonparametric regression, microarray data, gene network

1 Introduction

A Bayesian network is a powerful tool for modeling relations among a large number of random variables. Therefore the Bayesian network has received considerable attention from the studies of gene network estimation using microarray gene expression data. Imoto *et al.* [1, 2] proposed a Bayesian network and nonparametric regression model for capturing nonlinear relations between genes from the continuous gene expression data. However, a Bayesian network still has a problem that it cannot construct cyclic regulations, while real gene networks have cyclic regulations. For a solution of this problem, in this paper, we propose a dynamic Bayesian network and nonparametric regression model for estimating a gene network with cyclic regulations from time series microarray data. We also derive a criterion for selecting a network from Bayes approach. The effectiveness of our method is displayed though the analysis of the *Saccharomyces cerevisiae* gene expression data.

2 Method

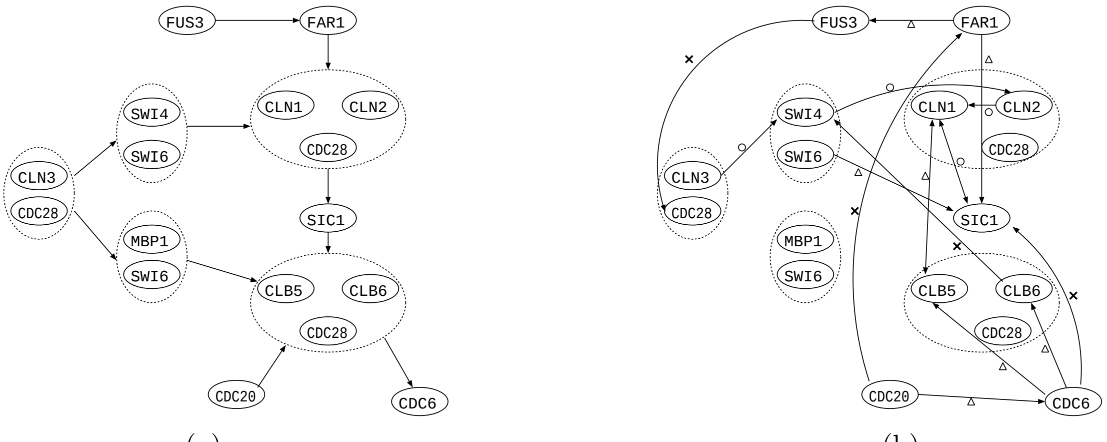
Let X be an $n \times p$ time series microarray data matrix, where n and p are the number of microarrays and genes, respectively. Under the first order Markov relation between the time points, the joint probability can then be decomposed as $P(X_{11}, \cdots, X_{np}) = P(\boldsymbol{X}_1)P(\boldsymbol{X}_2|\boldsymbol{X}_1) \times \cdots \times P(\boldsymbol{X}_n|\boldsymbol{X}_{n-1})$, where $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})^T$ is a random variable vector at time i. The conditional probability $P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1})$ can be decomposed as $P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1}) = P(X_{i1}|\boldsymbol{P}_{i-1,1}) \times \cdots \times P(X_{ip}|\boldsymbol{P}_{i-1,p})$, where $\boldsymbol{P}_{i-1,j}$ denotes the parents of jth gene at time $i-1$.

Using the nonparametric regression in order to model the relationship between a gene and its parents, we define a dynamic Bayesian network and nonparametric regression model by the density,

$$f(x_{11}, \cdots, x_{np}; \boldsymbol{\theta}_G) = f_1(\boldsymbol{x}_1) \prod_{j=1}^{p} \left[\prod_{i=2}^{n} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{ -\frac{(x_{ij} - \mu(\boldsymbol{p}_{i-1,j}))^2}{2\sigma_j^2} \right\} \right],$$

where $\boldsymbol{p}_{i-1,j} = (p_{i-1,1}^{(j)}, \cdots, p_{i-1,q_j}^{(j)})$ is a parents vector of jth gene, observed at time $i-1$.

When the network structure is given, we can construct a gene network by using the proposed model. However, the true gene network is still unknown, and we should guess the optimal network structure from the data. We derive a criterion for evaluating the network structure from Bayes approach. By using the Laplace approximation for integrals, the criterion, named BNRC$_{dynamic}$ can be expressed as



(a) (b)

Figure 1: Yeast cell cycle pathway compiled by KEGG. (a) Target, (b) Estimate.

$$\begin{aligned}\mathrm{BNRC}_{dynamic}(G) &= -2\log\left\{\pi_{prior}(G)\int f(x_{11},\cdots,x_{np};\theta_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G\right\}\\ &\approx -2\log\pi_{prior}(G) - r\log(2\pi/n) + \log|J_\lambda(\hat{\boldsymbol{\theta}}_G)| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}),\end{aligned}$$

where $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ and $\pi_{prior}(G)$ are the prior distribution of the parameter $\boldsymbol{\theta}_G$ and the prior probability of the network G, respectively, $\boldsymbol{\lambda}$ is the hyper parameter vector, r is the dimension of $\boldsymbol{\theta}_G$, $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}) = \log f(x_{11},\cdots,x_{np};\boldsymbol{\theta}_G)/n + \log\pi(\boldsymbol{\theta}_G|\lambda)/n$, $J_\lambda(\boldsymbol{\theta}_G) = -\partial^2\{l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})\}/\partial\boldsymbol{\theta}_G\partial\boldsymbol{\theta}_G^T$ and $\hat{\boldsymbol{\theta}}_G$ is the mode of $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})$. We can choose the optimal network such that the $\mathrm{BNRC}_{dynamic}$ is minimal.

3 Result

We apply the proposed method to the *Saccharomyces cerevisiae* cell cycle data collected by Spellman *et al.* [3]. The target network is a part of cell cycle pathway compiled by KEGG [4] and shown in Figure 1 (a). Figure 1 (b) is the estimated network based on the proposed method. In Figure 1 (b), we evaluate the estimated edges by three kinds of marks: Round is the correct edge, crisscross is the wrong edge and triangle represents the misdirection or skip.

References

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

[2] 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.

[3] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, 9:3273–3297, 1998.

[4] `http://www.genome.ad.jp/kegg/`