ESTIMATION OF GENETIC NETWORKS AND FUNCTIONAL STRUCTURES BETWEEN GENES BY USING BAYESIAN NETWORKS AND NONPARAMETRIC REGRESSION

SEIYA IMOTO, TAKAO GOTO and SATORU MIYANO
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

We propose a new method for constructing genetic network from gene expression data by using Bayesian networks. We use nonparametric regression for capturing nonlinear relationships between genes and derive a new criterion for choosing the network in general situations. In a theoretical sense, our proposed theory and methodology include previous methods based on Bayes approach. We applied the proposed method to the *S. cerevisiae* cell cycle data and showed the effectiveness of our method by comparing with previous methods.

1 Introduction

The microarray technology provides us enormous amount of valuable gene expression data. The analysis of the relationship among genes has drawn remarkable attention in the field of molecular biology and Bioinformatics. However, due to the cause of dimensionality and complexity of the data, it will be no easy task to find structures, which are buried in noise. To extract the effective information from microarray gene expression data, thus, theory and methodology are expected to be developed from a statistical point of view. Our purpose is to establish a new method for extracting the relationships among genes clearer.

Constructing genetic networks[3,4,5,12,13,19] is one of the hot topics in the analysis of the microarray gene expression data. Bayesian network is an attractive method for constructing genetic networks from a graph-theoretic approach. Friedman and Goldszmidt[12] proposed an interesting method for constructing genetic links by using Bayesian networks. They discretized the expression value and considered to fit the models based on multinomial distributions. However, a problem still remains to be solved in choosing the threshold value for discretizing not only by the experiments. The threshold value assuredly gives essential changes of the results and unsuitable threshold value leads to wrong results. On the other hand, recently, Friedman *et al.*[13] pointed out that discretizing is probably loosing the information. To use the expression data as continuous values, thus, they considered the use of Gaussian models based on linear regression. However this model can only detect linear dependencies and we cannot obtain sufficient results.

In this paper we propose a new method for constructing genetic networks

by using Bayesian networks. To capture not only linear dependencies but also nonlinear structures between genes, we use nonparametric regression models with Gaussian noise[11,15,22,23]. Nonparametric regression has been developed in order to explore the complex nonlinear form of the expected responses without the knowledge about the functional relationship in advance. Due to the new structure of the Bayesian networks, a suitable criterion is needed for evaluating models. We derive a new criterion from Bayesian statistics. By using proposed method, we will overcome the defects of previous methods and attain more effective information. In addition, our method includes the previous method[12] as a special case. The efficiency of the proposed method is shown by the Monte Carlo simulation method. We also demonstrate our proposed method through the analysis of the *S. cerevisiae* cell cycle data[21].

2 Bayesian Network and Nonparametric Regression

Let $\boldsymbol{X} = (X_1, X_2, ..., X_p)^T$ be a p-dimensional random vector and let G be a directed acyclic graph. Under the Bayesian network framework, we look upon a gene as a random variable and decompose the joint probability into the product of conditional probabilities, that is

$$P(X_1, X_2, ..., X_p) = P(X_1|\boldsymbol{P}_1)P(X_2|\boldsymbol{P}_2) \times \cdots \times P(X_p|\boldsymbol{P}_p), \tag{1}$$

where $\boldsymbol{P}_j = (P_1^{(j)}, P_2^{(j)}, ..., P_{q_j}^{(j)})^T$ is a q_j-dimensional vector of parent variables of X_j in the graph G.

Suppose that we have n observations $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ of the random vector $\boldsymbol{X}$ and the observations of $\boldsymbol{P}_j$ are denoted by $\boldsymbol{p}_{1j}, ..., \boldsymbol{p}_{nj}$, where $\boldsymbol{p}_{ij}$ is a q_j-dimensional vector with k-th element $p_{ik}^{(j)}$, for $k = 1, ..., q_j$. For example, let $\boldsymbol{X}_n$ be an $n \times p$ matrix, where $\boldsymbol{X}_n = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T = (\boldsymbol{x}_{(1)}, ..., \boldsymbol{x}_{(p)}) = (x_{ij})_{i=1,...,n;j=1,...,p}$, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$, $\boldsymbol{x}_{(j)} = (x_{1j}, ..., x_{nj})^T$ and $\boldsymbol{x}_i^T$ is the transpose of the vector $\boldsymbol{x}_i$. If X_1 has a parent vector $\boldsymbol{P}_1 = (X_2, X_3)^T$, we obtain $\boldsymbol{p}_{11} = (x_{12}, x_{13})^T, ..., \boldsymbol{p}_{n1} = (x_{n2}, x_{n3})^T$.

It is immediately found that the equation holds when we replace the probability measure P in (1) by densities

$$f(x_{i1}, x_{i2}, ..., x_{ip}) = f_1(x_{i1}|\boldsymbol{p}_{i1})f_2(x_{i2}|\boldsymbol{p}_{i2}) \times \cdots \times f_p(x_{ip}|\boldsymbol{p}_{ip}).$$

Then all we need to do is to consider how to construct the conditional densities $f_j(x_{ij}|\boldsymbol{p}_{ij})$ $(j = 1, ..., p)$.

In this paper, we use nonparametric regression models for capturing the relationship between x_{ij} and $\boldsymbol{p}_{ij} = (p_{i1}^{(j)}, ..., p_{iq_j}^{(j)})^T$ in the form

$$x_{ij} = m_1(p_{i1}^{(j)}) + m_2(p_{i2}^{(j)}) + \cdots + m_{q_j}(p_{iq_j}^{(j)}) + \varepsilon_{ij},\ i = 1, ..., n;\ j = 1, ..., p,$$

where m_k $(k = 1, ..., q_j)$ are smooth functions from R (a set of the real number) to R and ε_{ij} $(i = 1, ..., n)$ depend independently and normally on mean 0 and variance σ_j^2. For the function m_k, it is assumed that

$$m_k(p_{ik}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)}), \quad i = 1, ..., n;\ k = 1, ..., q_j, \tag{2}$$

where $\{b_{1k}^{(j)}, ..., b_{M_{jk}k}^{(j)}\}$ is a prescribed set of basis functions (such as Fourier series, polynomial bases, regression spline bases, B-spline bases, wavelet bases and so on), the coefficients $\gamma_{1k}^{(j)}, ..., \gamma_{M_{jk}k}^{(j)}$ are unknown parameters and M_{jk} is the number of basis functions.

Then a nonparametric regression model can be written as a probability density function in the form

$$f_j(x_{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\gamma}_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{\{x_{ij} - \sum_{k=1}^{q_j} \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)})\}^2}{2\sigma_j^2} \right], \tag{3}$$

where $\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j1}^T, ..., \boldsymbol{\gamma}_{jq_j}^T)^T$ is a parameter vector with $\boldsymbol{\gamma}_{jk} = (\gamma_{1k}^{(j)}, ..., \gamma_{M_{jk}k}^{(j)})^T$. If a variable X_j has no parent variables, we consider the model based on the normal distributions with mean μ_j and variance σ_j^2.

Finally we have a Bayesian network model based on the nonparametric regression model with Gaussian noise

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) = \prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\theta}_j), \quad i = 1, ..., n,$$

where $\boldsymbol{\theta}_G = (\boldsymbol{\theta}_1^T, ..., \boldsymbol{\theta}_p^T)^T$ is a parameter vector included in the graph G and $\boldsymbol{\theta}_j$ is a parameter vector in the model f_j, i.e., $\boldsymbol{\theta}_j = (\boldsymbol{\gamma}_j^T, \sigma_j^2)^T$ or $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2)^T$.

3 Proposed criterion for choosing graph

Let $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ be the prior distribution on the unknown parameter vector $\boldsymbol{\theta}_G$ with hyper parameter vector $\boldsymbol{\lambda}$ and let $\log \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(n)$. The marginal probability of the data $\boldsymbol{X}_n$ is obtained by integrating over the parameter space, and we choose a graph G with the largest posterior probability

$$\pi_G \int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) d\boldsymbol{\theta}_G, \tag{4}$$

where π_G is a prior probability of G. Friedman and Goldszmidt[12] considered the multinomial distribution as the Bayesian network model $f(\boldsymbol{x}_i;\boldsymbol{\theta}_G)$, and also supposed the Dirichlet prior on the parameter $\boldsymbol{\theta}_G$. In this case, the Dirichlet prior is the conjugate prior and the posterior distribution belongs to the same class of distribution. Then a closed form solution of the integration in (4) is obtained, and they called it BDe score for choosing graph[6,16]. Recall that the BDe score is confined to the multinomial model, and we propose a criterion for choosing graph in more general and various situations.

The essential problem of constructing criteria based on (4) is how to compute the integration. While some methods can be considered for computing the integration such as Markov chain Mote Carlo, we use the Laplace approximation for integrals[7,17,24], because it is not necessary to consider the conjugate prior. The Laplace approximation to the marginal probability of $\boldsymbol{X}_n$ is

$$\begin{aligned}\int \prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G &= \int \exp\{nl_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)\}d\boldsymbol{\theta}_G \\ &= \frac{(2\pi/n)^{r/2}}{|J_\lambda(\hat{\boldsymbol{\theta}}_G)|^{1/2}}\exp\{nl_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}_n)\}\{1+O_p(n^{-1})\},\end{aligned}$$

where r is the dimension of $\boldsymbol{\theta}_G$,

$$\begin{aligned}l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n) &= \frac{1}{n}\sum_{i=1}^{n}\log f(\boldsymbol{x}_i;\boldsymbol{\theta}_G) + \frac{1}{n}\log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}), \\ J_\lambda(\boldsymbol{\theta}_G) &= -\partial^2\{l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)\}/\partial\boldsymbol{\theta}_G\partial\boldsymbol{\theta}_G^T\end{aligned}$$

and $\hat{\boldsymbol{\theta}}_G$ is the mode of $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)$. Then we have a criterion, BNRC, for selecting graph

$$\begin{aligned}\mathrm{BNRC}(G) &= -2\log\left\{\pi_G\int\prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G\right\} \\ &= -2\log\pi_G - r\log(2\pi/n) + \log|J_\lambda(\hat{\boldsymbol{\theta}}_G)| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}_n). \qquad (5)\end{aligned}$$

The optimal graph is chosen such that the criterion BNRC (5) is minimal.

This criterion is derived under $\log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(n)$. If $\log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(1)$, the mode $\hat{\boldsymbol{\theta}}_G$ is equivalent to the maximum likelihood estimate, MLE, and the criterion is resulted in Bayesian information criterion, known as BIC[20] by removing the higher order terms $O(n^{-j})$ $(j \geq 0)$. Konishi[18] provided a general framework for constructing model selection criteria based on the Kullback-Leibler information and Bayes approach.

It is assumed that the prior density $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ is decomposed into the product of the prior densities on $\boldsymbol{\theta}_j$, $\pi_G(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \pi_1(\boldsymbol{\theta}_1|\boldsymbol{\lambda}_1)\times\cdots\times\pi_p(\boldsymbol{\theta}_p|\boldsymbol{\lambda}_p)$. Hence $l_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}_n)$ and $\log|J_\lambda(\hat{\boldsymbol{\theta}}_G)|$ in (5) are, respectively,

$$\sum_{j=1}^{p} l_\lambda^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}_n) \quad \text{and} \quad \sum_{j=1}^{p}\log\left|-\frac{\partial^2 l_\lambda^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}_n)}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T}\right|,$$

where

$$l_\lambda^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\log f_j(x_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\theta}_j) + \frac{1}{n}\log\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j). \tag{6}$$

Thus the BNRC (5) can be obtained by the local scores of graph as follows: We define the local BNRC for the j-th variable X_j by

$$\mathrm{BNRC}_j = -2\log\left\{\pi_{L_j}\int\prod_{i=1}^{n} f_j(x_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)d\boldsymbol{\theta}_j\right\}, \tag{7}$$

where π_{L_j} is a prior probability of the local structure associated with X_j. We also apply Laplace's method to the BNRC_j and the BNRC is obtained by

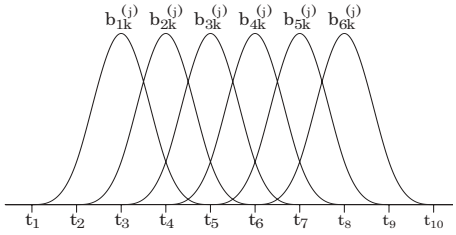
$$\mathrm{BNRC} = -2\log\pi_G + \sum_{j=1}^{p}\{\mathrm{BNRC}_j + 2\log\pi_{L_j}\}.$$

Notice that the final graph is selected as a minimizer of the BNRC and it is not necessary minimize each local score BNRC_j, because the graph is constructed as acyclic.

4 Estimating graph and related structures by using BNRC

In this section we express our method in more concrete terms. The key idea of our proposed method is the use of the nonparametric regression and the new criterion for choosing graph from Bayesian statistics.

As for nonparametric regression in Section 2, we use the B-splines[8] as the basis functions in (2). Figure 1 is an example of B-splines of degree 3 with equidistance knots $t_1, ..., t_{10}$. We place the knots dividing the domain $[\min_i(p_{ik}^{(j)}), \max_i(p_{ik}^{(j)})]$ into $M_{jk}-3$ equidistance interval[10] and set M_{jk} B-splines of degree 3.



Figrure 1: Example of 6 B-splines of degree 3. $t_1, ..., t_{10}$ are called knots. These knots are equally spaced.

We assume that the prior distribution on the parameter vector $\boldsymbol{\theta}_j$ is

$$\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) = \prod_{k=1}^{q_j} \pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk}).$$

Each prior distribution $\pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk})$ is a singular M_{jk} variate normal distribution given by

$$\pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk}) = \left(\frac{2\pi}{n\lambda_{jk}}\right)^{-(M_{jk}-2)/2} |K_{jk}|_+^{1/2} \exp\left(-\frac{n\lambda_{jk}}{2}\boldsymbol{\gamma}_{jk}^T K_{jk}\boldsymbol{\gamma}_{jk}\right), \quad (8)$$

where λ_{jk} is a hyper parameter, K_{jk} is an $M_{jk} \times M_{jk}$ matrix, $\boldsymbol{\gamma}_{jk}^T K_{jk}\boldsymbol{\gamma}_{jk} = \sum_{l=3}^{M_{jk}} (\gamma_{lk}^{(j)} - 2\gamma_{l-1,k}^{(j)} + \gamma_{l-2,k}^{(j)})^2$ and $|K_{jk}|_+$ is the product of $M_{jk}-2$ nonzero eigenvalues of K_{jk}. The score BNRC_j (7) can be obtained as

$$\begin{aligned}\mathrm{BNRC}_j &= -2\log \pi_{L_j} - 2\sum_{i=1}^{n} \log f_j(x_{ij}|\boldsymbol{p}_{ij}; \hat{\boldsymbol{\theta}}_j) - 2\sum_{k=1}^{q_j} \log \pi_k(\hat{\boldsymbol{\gamma}}_{jk}|\lambda_{jk}) \\ &+ \log\left|-\frac{\partial^2 l_\lambda^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}_n)}{\partial\boldsymbol{\theta}_j\boldsymbol{\theta}_j^T}\right| - (\sum_{k=1}^{q_j} M_{jk} + 1)\log(2\pi n^{-1}), \end{aligned} \quad (9)$$

where $\hat{\boldsymbol{\theta}}_j = (\hat{\boldsymbol{\gamma}}_j^T, \hat{\sigma}_j^2)^T$ is a mode of $l_\lambda^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}_n)$ defined in (6) for fixed λ_{jk}. For computational aspect, we approximate the logarithm of the determinant of the Hessian matrix in (9) by

$$\sum_{k=1}^{q_j}\{\log|B_{jk}^T B_{jk} + n\hat{\sigma}_j^2\lambda_{jk}K_{jk}| - M_{jk}\log(n\hat{\sigma}_j^2)\} - \log(2\hat{\sigma}_j^4),$$

where B_{jk} is an $n \times M_{jk}$ matrix defined by $B_{jk} = (\boldsymbol{b}_{jk}(p_{1k}^{(j)}), ..., \boldsymbol{b}_{jk}(p_{nk}^{(j)}))^T$ with $\boldsymbol{b}_{jk}(p_{ik}^{(j)}) = (b_{1k}^{(j)}(p_{ik}^{(j)}), ..., b_{M_{jkk}}^{(j)}(p_{ik}^{(j)}))^T$. Hence combining (3), (8) and (9), BNRC_j is resulted in

$$\begin{aligned}\mathrm{BNRC}_j &= C_j + (n - 2q_j - 2)\log\hat{\sigma}_j^2 \\ &+ \sum_{k=1}^{q_j}\left\{\frac{n\beta_{jk}}{\hat{\sigma}_j^2}\hat{\boldsymbol{\gamma}}_{jk}^T K_{jk}\hat{\boldsymbol{\gamma}}_{jk} + \log|\Lambda_{jk}| - (M_{jk}-2)\log\beta_{jk}\right\},\end{aligned}$$

where $\beta_{jk} = \sigma_j^2\lambda_{jk}$ is a hyper parameter,

$$\begin{aligned}C_j &= -2\log\pi_{L_j} + (n + \bar{M}_{j\cdot} - 2q_j)\log(2\pi) + n - \log 2 \\ &\quad -2(\bar{M}_{j\cdot} - q_j)\log n - \sum_{k=1}^{q_j}\log|K_{jk}|_+,\end{aligned}$$

$$\Lambda_{jk} \quad = \quad B_{jk}^T B_{jk} + n\beta_{jk} K_{jk}, \qquad \bar{M}_{j\cdot} = \sum_{k=1}^{q_j} M_{jk}.$$

By using the backfitting algorithm[15], the modes $\hat{\boldsymbol{\gamma}}_{jk}$ $(k = 1, ..., q_j)$ can be obtained when the values of β_{jk} are given. The backfitting algorithm can be expressed as follow:
Step 1 *Initialize:* $\boldsymbol{\gamma}_{jk} = \mathbf{0}, \quad k = 1, ..., q_j.$
Step 2 *Cycle:* $k = 1, ..., q_j, 1, ..., q_j, 1, ...$

$$\boldsymbol{\gamma}_{jk} = (B_{jk}^T B_{jk} + n\beta_{jk} K_{jk})^{-1} B_{jk}^T (\boldsymbol{x}_{(j)} - \sum_{k' \neq k} B_{jk'} \boldsymbol{\gamma}_{jk'}).$$

Step 3 *Continue Step 2 until a suitable convergence criterion is satisfied.*
The mode $\hat{\sigma}_j^2$ is given by $\hat{\sigma}_j^2 = ||\boldsymbol{x}_{(j)} - \sum_{k=1}^{q_j} B_{jk} \hat{\boldsymbol{\gamma}}_{jk}||^2 / n$.

In attention, the modes $\hat{\boldsymbol{\gamma}}_{jk}$ and $\hat{\sigma}_j^2$ depend on the hyper parameters β_{jk} and we have to choose the optimal values of β_{jk}. In the context of our method, it is natural that the optimal β_{jk} are chosen as the minimizers of BNRC_j.

Recall that the B-splines coefficients vectors $\boldsymbol{\gamma}_{jk}$ are estimated by maximizing (6). The modes of (6) are the same as the penalized likelihood estimates[22,25] and we can look upon the hyper parameters λ_{jk} or β_{jk} as the smoothing parameters in penalized likelihood. Hence, the hyper parameters play an important role for fitting the curve to the data.

5 Computational experiments

Monte Carlo simulation

Before analyzing the real data, we used the Monte Carlo simulation method to examine the effectiveness of our method. The data were generated from an artificial graph and structures between variables (Figure 2).

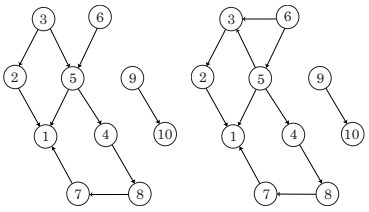


Figure 2: Monte Carlo simulation. (Left) true, (Right) estimate.

$$X_1 = X_2^2 + 2\sin(X_5) - 2X_7 + \varepsilon_1$$
$$X_2 = \{1 + \exp(-4X_3)\}^{-1} + \varepsilon_2$$
$$X_3 = \varepsilon_3, \ X_6 = \varepsilon_6, \ X_9 = \varepsilon_9$$
$$X_4 = X_5^2/3 + \varepsilon_4, \ X_5 = X_3 - X_6^2 + \varepsilon_5$$
$$X_7 = \begin{cases} -1 + \varepsilon_7, & X_8 \leq -0.5 \\ X_8 + \varepsilon_7, & -0.5 < X_8 \leq 0.5 \\ 1 + \varepsilon_7, & 0.5 < X_8 \end{cases}$$
$$X_8 = \exp(-X_4 - 1)/2 + \varepsilon_8$$
$$X_{10} = \cos(X_9) + \varepsilon_{10}.$$

The results from this Monte Carlo simulation can be summarized as follows: Proposed criterion BNRC can detect linear and nonlinear structures of the data very well. But the BNRC has a tendency toward overgrowth of graph. We then consider the use of Akaike's information criterion known as AIC[1,2] and use both methods. AIC was originally introduced as a criterion for evaluating models estimated by maximum likelihood method. But the estimate by our method is the same as the maximum penalized likelihood estimates and is not MLE. In this case, the modified version of AIC[10] is given by

$$\mathrm{AIC} = -2\sum_{i=1}^{n} \log f_j(x_{ij}|\boldsymbol{p}_{ij}; \hat{\boldsymbol{\gamma}}_j, \hat{\sigma}_j^2) + 2(\sum_{k=1}^{q_j} \mathrm{tr}S_{jk} + 1),$$

where $S_{jk} = B_{jk}(B_{jk}^T B_{jk} + n\hat{\beta}_{jk} K_{jk})^{-1} B_{jk}^T$. The trace of S_{jk} shows the degree of freedom of the fitted curve and is a great help. That is to say, if $\mathrm{tr}S_{jk}$ is nearly 2, the dependency is looked upon linear. We use both BNRC and AIC for decision whether we add up to a parent variable. By using this method, the estimated graph and structures are close to the true model.

Analysis of cell cycle data

We analyze the *S.cerevisiae* cell cycle data discussed by Spellman *et al.*[21] and Friedman *et al.*[13]. The data were collected from 800 genes with 77 experiments.

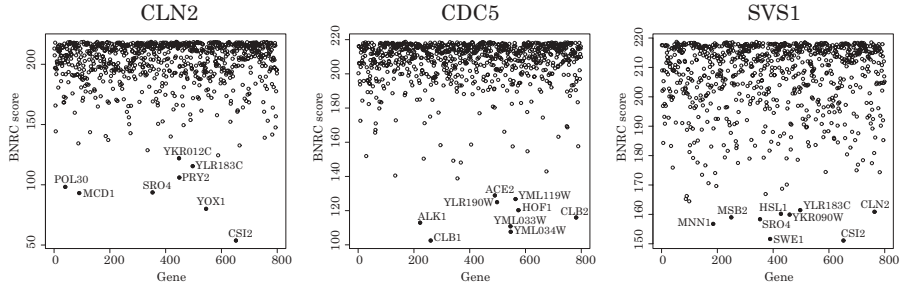


Figure 3: BNRC scores for CNL2, CDC5 and SVS1.

We set the prior probability π_G constant, because we have no reason why the large graph is unacceptable and no information about the size of the true graph. The nonparametric regressors are constructed with 20 B-splines. In fact, the number of B-splines is also a parameter. However, we use somewhat large number of B-splines, the hyper parameters control the smoothness of fitted curve and we cannot visually find differences among fitted curves corresponding to various number of B-splines.

The results of the analysis from the cell cycle data can be summarized as follows: Figure 3 shows BNRC scores when we predict CLN2, CDC5 and SVS1 by one gene. The genes, which give smaller BNRC scores, give a better expression to the target gene. We can observe that which gene is associated with the target gene and we find the set of genes which strongly depend on the target gene. In fact, we can construct a brief network by using this information. We can look upon the optimal graph as a revised version of the brief network by choosing the parent genes and holding the assumption of acyclic. We note that if there is a linear dependency between genes, the score BNRC is also good when the parent-child relationship is reversed. Therefore, the directions of the causal associations in the graph are not strict especially when the dependency is almost linear. Our result basically advocates the result of Friedman *et al.*[13], but, of cause, there are different points in parts. There are some genes that mediate Friedman *et al.*'s result, such as MCD1, CSI2, YOX1 and so on. A large number of the relationships between genes are nearly linear. But we could find some nonlinear dependencies which linear models hardly find. Figure 5 shows the estimated graph associated with genes which were classified their processes into cell cycle and their neighborhood. Here, we omit some branches in Figure 5, but important information is almost shown. As for the networks given by us and Friedman *et al.*[13], we confirmed parent-children relationships and observed that both two networks are similar to each other. Especially, our network includes typical relationships which were reported by Friedman *et al.*[13]. As for the differences between two networks, we paid attention to the parent genes of SVS1. Friedman *et al.*[13] employed CLN2 and CDC5 as the parent genes of SVS1. On the one hand, our result gives CSI2 and YKR090W for SVS1. We check up on the difference of these two results. In the sense of BNRC and AIC, our candidate parent genes are more appropriate than Friedman *et al.*[13]'s. The reason might be the effect of discretizing, because our model suitably fits to both cases in Figure 4. We notice that the range of the fitted curve in Figure 4 (b) is much smaller than other curves. All in all, we conclude that CDC5 gives just weak effects to SVS1 compared with other genes from Spellman *et al.*[21]'s data (see also Figure 3). In fact, as the parent gene of SVS1, the order of BNRC score of CDC5 is 247th. Considering the circumstances mentioned above, our method can provide us valuable information in understandable and useful form.

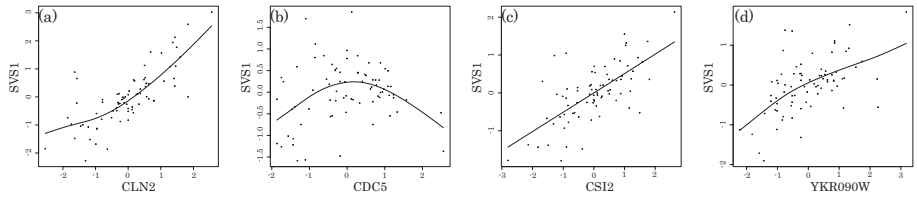


Figure 4: Cell cycle data and smoothed estimates.
(a) and (b) Friedman *et al.*[13], BNRC = 160.45, AIC=167.96;
(c) and (d) Proposed method, BNRC = 135.27, AIC=140.16.

6 Discussion

We proposed the new method for estimating genetic networks from microarray gene expression data by using Bayesian network and nonparametric regression. We derived a new criterion for choosing graph theoretically, and represented its effectiveness through the Monte Carlo simulations and the analysis of the cell cycle data. The advantages of our method are mainly as follows: We can use the expression data as continuous values. Not only linear dependencies, we can also detect nonlinear structures and can visualize their functional structures being easily understandable. Fully automatic search can accomplish the creation of optimal graph.

We also pointed out that Friedman *et al.*[13]'s method remained the unknown parameters such as threshold value for discretizing and hyper parameters in the Dirichlet priors which selected by trial and error. These parameters were not optimized in a narrow sense. On the other hand, our proposed method can automatically and appropriately estimate any parameters based on proposed criterion which has a sounder theoretical basis. Besides, our method includes Friedman *et al.*[13]'s as a special case.

We consider the following problems as our future works: (1) We used the statistical models based on Gaussian distribution. However, we derive the criterion BNRC in more general situations. In fact, we can construct the graph selection criterion based on other statistical models. (2) It is a possible case that the outliers cause strange results. Thus, the development of the robust methods and the technique for detecting the outliers are important problems. (3) The intensities of the unions are probably measured by using bootstrap method[9]. We would like to investigate these problems in a future paper.

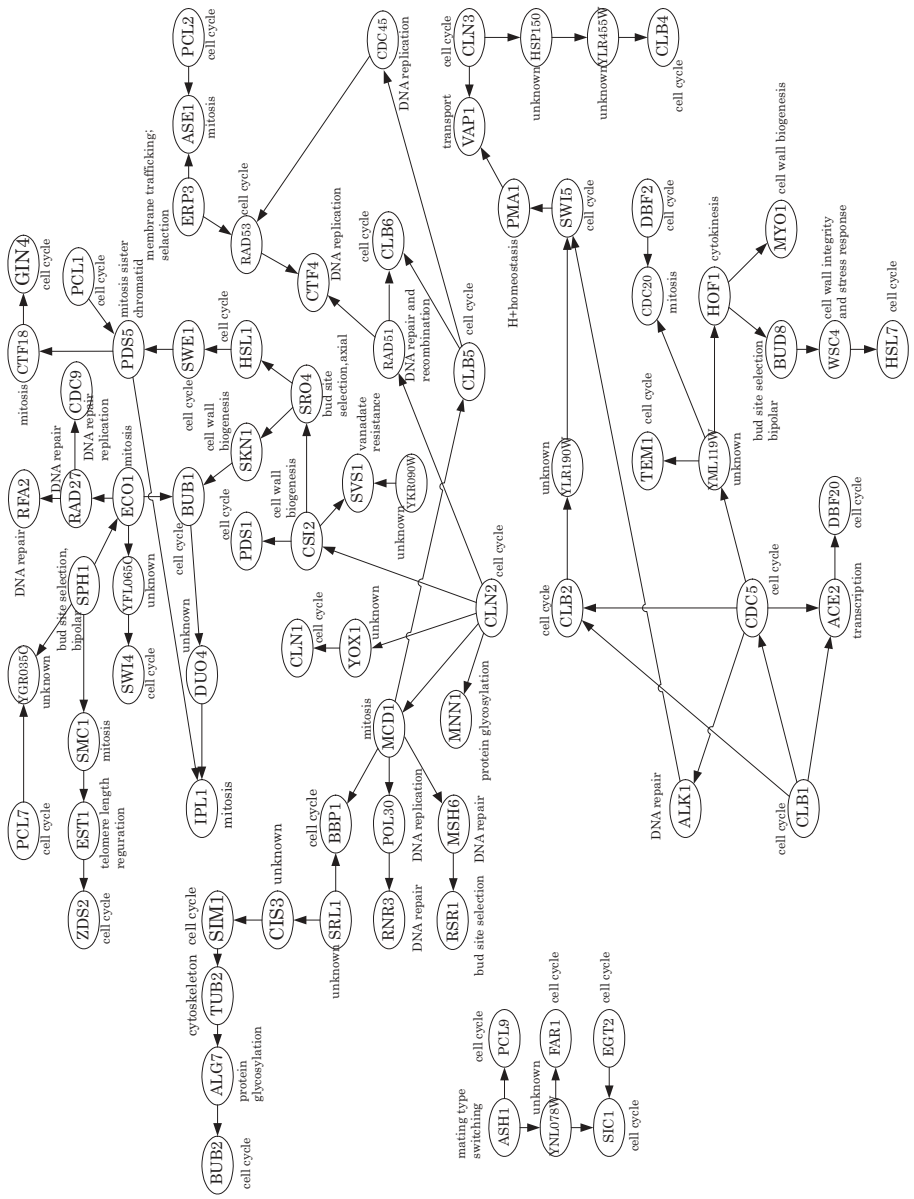


Figure 5: Cellcycle data result.

References

1. H. Akaike, *in* Petrov,B.N. and Csaki,F. eds., *2nd Inter. Symp. on Information Theory*, Akademiai Kiado, Budapest, 267 (1973).
2. H. Akaike, *IEEE Trans. Autom. Contr.*, **AC-19**, 716 (1974).
3. T. Akutsu, S. Miyano and S. Kuhara, *Pacific Symposium on Biocomputing*, 17, (1999).
4. T. Akutsu, S. Miyano and S. Kuhara, *Bioinformatics*, **16**, 727 (2000).
5. T. Akutsu, S. Miyano and S. Kuhara, *J. Comp. Biol.,* **7**, 331 (2000).
6. G. F. Cooper and E. Herskovits, *Machine Learning*, **9**, 309 (1992).
7. A. C. Davison, *Biometrika*, **73**, 323 (1986).
8. C. de Boor, *A Practical Guide to Splines*. Springer, Berlin. (1978).
9. B. Efron, *Ann. Stat.*, **7**, 1 (1979).
10. P. H. C. Eilers and B. Marx, *Statistical Science*, **11**, 89 (1996).
11. R. L. Eubank, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York. (1988).
12. N. Friedman and M. Goldszmidt, *in* M. I. Jordan ed., *Learning in Graphical Models*, Kluwer Academic Publisher. 421 (1998).
13. N. Friedman, M. Linial, I. Nachman and D. Pe'er, *J. Comp. Biol.,* **7**, 601 (2000).
14. P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London. (1994).
15. T. Hastie and R. Tibshirani, *Generalized Additive Models*. Chapman & Hall, London. (1990).
16. D. Heckerman, D. Geiger and D. M. Chickering, *Machine Learning*, **20**, 274 (1995).
17. D. Heckerman, *in* M. I. Jordan ed., *Lerning in Graphical Models* 301, Kluwer Academic Publisher. (1998).
18. S. Konishi, (in Japanese), *Sugaku*, **52**, 128 (2000).
19. D. Pe'er, A. Regev, G. Elidan and N. Friedman, *Bioinformatics*, **17** *Suppl.1*, 215 (ISMB 2001).
20. G. Schwarz, *Ann. Stat.*, **6**, 461 (1978).
21. P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein and B. Futcher, *Mol. Biol. Cell*, **9**, 3273 (1998).
22. B. W. Silverman, *J. R. Stat. Soc. Series* B, **47**, 1 (1985).
23. J. S. Simonoff, *Smoothing Methods in Statistics*. Springer, New York. (1996).
24. L. Tinerey and J. B. Kadane, *J. Amer. Statist. Assoc.*, **81**, 82 (1986).
25. G. Wahba, *J. R. Stat. Soc. Series* B, **40**, 364 (1978).