Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems 75 (2004) 57–65
BioSystems
www.elsevier.com/locate/biosystems

Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data

Sunyong Kim, Seiya Imoto, Satoru Miyano*

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

Abstract

We propose a dynamic Bayesian network and nonparametric regression model for constructing a gene network from time series microarray gene expression data. The proposed method can overcome a shortcoming of the Bayesian network model in the sense of the construction of cyclic regulations. The proposed method can analyze the microarray data as a continuous data and can capture even nonlinear relations among genes. It can be expected that this model will give a deeper insight into complicated biological systems. We also derive a new criterion for evaluating an estimated network from Bayes approach. We conduct Monte Carlo experiments to examine the effectiviness of the proposed method. We also demonstrate the proposed method through the analysis of the *Saccharomyces cerevisiae* gene expression data.

Keywords: Microarrays; Gene networks; Dynamic Bayesian networks

1. Introduction

The development of microarray technology provides us a huge amount of gene expression data and a new perspective of the analysis of whole genome mechanism. The estimation of a gene network from cDNA microarray gene expression data becomes one of the important topics in the field of bioinformatics and can be viewed as the first step of systems biology.

The use of the Bayesian network model (Friedman et al., 2000; Imoto et al., 2002a,b; Pe'er et al., 2001) for estimating a gene network from cDNA microarray gene expression data has received considerable attention and many successful investigations have been reported. However, a shortcoming of the Bayesian network model is that this model cannot construct cyclic networks, while a real gene regulation mechanism has cyclic regulations. Recently, the dynamic Bayesian network model (Bilmes, 2000; Friedman et al., 1998; Ong et al., 2002; Someren et al., 2002) has been proposed for constructing a gene network with cyclic regulations. The dynamic Bayesian network is based on time series data. Usually the data is discretized into several classes. Therefore, the resulting network of the dynamic Bayesian network model depends strongly on the setting of the thresholds for discretization, and, unfortunately, the discretization leads to information loss. Imoto et al. (2002a,b) proposed the network estimation method based on the Bayesian network and the nonparametric regression for a solution to avoid the discretization and for capturing nonlinear relations among genes.

In this paper, we extend the Bayesian network and nonparametric regression model to the dynamic Bayesian network model, which can construct cyclic

* Corresponding author.
E-mail address: miyano@ims.u-tokyo.ac.jp (S. Miyano).

0303-2647/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2004.03.004

regulations when we have time series gene expression data. We can include the information of time delay into the proposed model naturally and the model can extract even nonlinear relations among genes automatically. For constructing a gene network with cyclic regulations based on time series gene expression data, an ordinary differential equation model (Chen et al., 1999; De Hoon et al., 2003) is an alternative method. However, most implementations using this model are based only on linear systems. They are probably unsuitable for capturing complex phenomena. We derive a new criterion for choosing an optimal network from the Bayesian statistical point of view (see Berger, 1985). The proposed criterion can optimize the network structure, which gives the best representation of the gene interactions described by the data with noise. The efficiency of the proposed method are shown through the analysis of the *Saccharomyces cerevisiae* gene expression data.

2. Dynamic Bayesian network and nonparametric regression

Let $\boldsymbol{X}$ be $n \times p$ microarray gene expression data matrix, where n and p are the numbers of microarrays and genes, respectively. In the context of Bayesian networks, a gene is considered as a random variable. When we model a gene network by using statistical models described by the density or probability function, the statistical model should include p random variables. However, we have only n samples and n is usually much smaller than p. In such case, the inference of the model is quite difficult or sometimes impossible, because the model has many parameters and the number of samples is not enough for estimating the parameters. A Bayesian network model has been advocated in such modeling.

While Bayesian networks are very effective for analyzing microarray data, they stand only when there are no cyclic dependencies. Dynamic Bayesian networks overcome this problem (see Fig. 1). In the context of the dynamic Bayesian networks, we consider the time series data and the ith row vector $\boldsymbol{x}_i$ of $\boldsymbol{X}$ corresponds to the states of p genes at time i. As for the time dependency, we consider the first order Markov relation described in Fig. 2. Under this condition, the joint probability can be decomposed as follows:

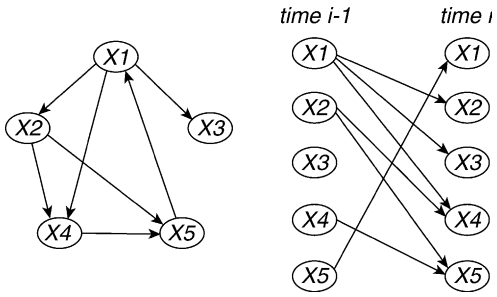


Fig. 1. Example of a network containing a cyclic regulation. The network (left) contains a cycle $X_1 \to X_2 \to X_4 \to X_5 \to X_1$. A Bayesian network model cannot treat such a network. On the other hand, the dynamic Bayesian network can construct a cyclic regulation by dividing states of a gene by time points (right).

$$P(X_{11}, \ldots, X_{np}) = P(\boldsymbol{X}_1)P(\boldsymbol{X}_2|\boldsymbol{X}_1) \times \cdots \times P(\boldsymbol{X}_n|\boldsymbol{X}_{n-1}), \quad (1)$$

where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ is a random variable vector of p genes at time i.

For each time slice, we construct a network representing gene regulations. As is shown in Fig. 2, we assume the network structure is stable through all time points. Taking these gene regulations, the conditional probability $P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1})$ can also be decomposed into the product of conditional probabilities of each gene given its parent genes, of the form

$$P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1}) = P(X_{i1}|\boldsymbol{P}_{i-1,1}) \times \cdots \times P(X_{ip}|\boldsymbol{P}_{i-1,p}), \quad (2)$$

where $\boldsymbol{P}_{i-1,j}$ is the state vector of the parent genes of jth gene at time $i-1$.

The Eqs. (1) and (2) hold when we use the density function instead of the probability measure. From Eq. (1), we have

$$f(x_{11}, \ldots, x_{np}) = f_1(\boldsymbol{x}_1)f_2(\boldsymbol{x}_2|\boldsymbol{x}_1) \times \cdots \times f_n(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}). \quad (3)$$

Suppose that $\boldsymbol{p}_{i-1,j} = (p^{(j)}_{i-1,1}, \ldots, p^{(j)}_{i-1,q_j})^{\mathrm{T}}$ is a q_j-dimensional observation vector of parent genes of jth gene at time $i-1$. The Eq. (2) can be rewritten as

$$f_i(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = g_1(x_{i1}|\boldsymbol{p}_{i-1,1}) \times \cdots \times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}). \quad (4)$$

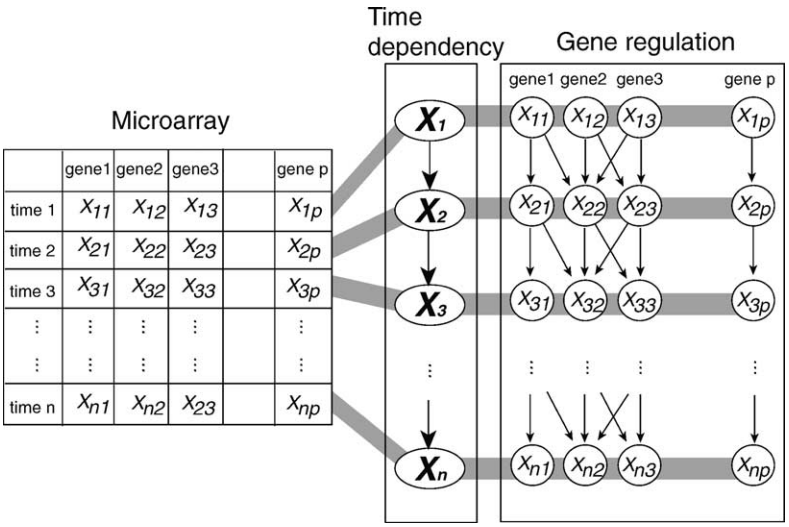


Fig. 2. Graphical view of a dynamic Bayesian network model.

By substituting (4) into (3), we have a dynamic Bayesian network model described by densities

$$\begin{aligned} f(x_{11},\ldots,x_{np}) &= f_1(\boldsymbol{x}_1) f_2(\boldsymbol{x}_2|\boldsymbol{x}_1) \times \cdots \times f_n(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \\ &= f_1(\boldsymbol{x}_1) \prod_{i=2}^{n} g_1(x_{i1}|\boldsymbol{p}_{i-1,1}) \cdots g_p(x_{ip}|\boldsymbol{p}_{i-1,p}) \\ &= f_1(\boldsymbol{x}_1) \prod_{j=1}^{p} \left\{ \prod_{i=2}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j}) \right\}. \end{aligned}$$

Hence, a crucial problem for modeling a gene network based on the dynamic Bayesian network is how to construct the conditional densities $g_j(x_{ij}|\boldsymbol{p}_{i-1,j})$. To construct this density function, we assume a nonparametric additive regression model with Gaussian noise,

$$x_{ij} = m_{j1}(p_{i-1,1}^{(j)}) + \cdots + m_{jq_j}(p_{i-1,q_j}^{(j)}) + \varepsilon_{ij},$$

where ε_{ij} depends independently and normally on mean 0 and variance σ_j^2. That is, $g_j(x_{ij}|\boldsymbol{p}_{i-1,j})$ is a density of Gaussian distribution. Here $m_{jk}(\cdot)$ is a smooth function from $\Re$ to $\Re$ and can be expressed by using the linear combination of basis functions

$$m_{jk}(p_{i-1,k}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{i-1,k}^{(j)}), \quad k = 1,\ldots,q_j,$$

where $\gamma_{1k}^{(j)},\ldots,\gamma_{M_{jk}k}^{(j)}$ are coefficient parameters and $\{b_{1k}^{(j)}(\cdot),\ldots,b_{M_{jk}k}^{(j)}(\cdot)\}$ is the prescribed set of basis functions. Then we define a dynamic Bayesian network and nonparametric regression model of the form

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

where $\boldsymbol{\theta}_{\mathrm{G}}$ is the parameter vector included in the model and $\mu(\boldsymbol{p}_{i-1,j}) = m_{j1}(p_{i-1,1}^{(j)}) + \cdots + m_{jq_j}(p_{i-1,q_j}^{(j)})$. When jth gene has no parent genes, $\mu(\boldsymbol{p}_{i-1,j})$ is resulted in the constant μ_j.

We assume $f_1(\boldsymbol{x}_1) = g_1(x_{11}) \times \cdots \times g_1(x_{1p})$ and the joint density $f(x_{11},\ldots,x_{np};\boldsymbol{\theta}_{\mathrm{G}})$ can then be rewritten as

$$\begin{aligned} f(x_{11},\ldots,x_{np};\boldsymbol{\theta}_{\mathrm{G}}) &= \prod_{j=1}^{p} \left\{ g_1(x_{1j}) \prod_{i=2}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j) \right\} \\ &= \prod_{j=1}^{p} \prod_{i=1}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j), \end{aligned} \tag{5}$$

where $\boldsymbol{p}_{0j} = \emptyset$. Thus, $g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j)$ represents the local structure of jth gene and its parent genes.

3. Derivation of a criterion for selecting networks

The dynamic Bayesian network and nonparametric regression model introduced in the previous section can be constructed when we fix the network structure and can be estimated by a suitable procedure. However, gene networks are generally unknown and we should estimate optimal networks based on the data. This problem can be viewed as a statistical model selection problem (see e.g., Akaike, 1973; Burnham and Anderson, 1998; Konishi, 1999; Konishi and Kitagawa, 1996). We solve this problem from the Bayesian statistical approach and derive a criterion for evaluating the goodness of the dynamic Bayesian network and nonparametric regression model.

Here we focus on a posterior probability of a network G since the optimal network is considered to maximize it. Let $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ be a prior distribution on the parameter $\boldsymbol{\theta}_G$ in the dynamic Bayesian network and nonparametric regression model and let $\log \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(n)$. The marginal likelihood can be represented as

$$\int f(x_{11}, \ldots, x_{np}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta}_G.$$

Thus, when the data is given, the posterior probability of the network G is

$$\pi_{\mathrm{post}}(G|\boldsymbol{X}) = \frac{\pi(G)\int f(x_{11}, \ldots, x_{np}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta}_G}{\sum_G \left\{\pi(G)\int f(x_{11}, \ldots, x_{np}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta}_G\right\}}, \tag{6}$$

where $\pi(G)$ is the prior probability of the network G. The denominator of (6) does not relate to model evaluation. Therefore, the evaluation of the network depends on the magnitude of numerator. Hence, we can choose an optimal network as the maximizer of

$$\pi(G)\int f(x_{11}, \ldots, x_{np}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta}_G.$$

It is clear that the essential point for constructing a network selection criterion is how to compute the high dimensional integral. Imoto et al. (2002a,b) used the Laplace approximation for integrals (see also Konishi et al., 2004; Tinerey and Kadane, 1986; Davison, 1986) and we can apply this technique to the dynamic Bayesian network and nonparametric regression model directly. Hence, we have a criterion, named $\mathrm{BNRC}_{\mathrm{dynamic}}$, of the form

$$\mathrm{BNRC}_{\mathrm{dynamic}}(G) = -2\log\left\{\pi(G)\int f(x_{11}, \ldots, x_{np}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\theta}_G\right\} \tag{7}$$

$$\approx -2\log\pi(G) - r\log\left(\frac{2\pi}{n}\right) + \log|J_\lambda(\hat{\boldsymbol{\theta}}_G)| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}), \tag{8}$$

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

$$l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}) = \log f(x_{11}, \ldots, x_{np}; \boldsymbol{\theta})/n + \log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})/n,$$

$$J_\lambda(\boldsymbol{\theta}_G) = -\frac{\partial^2\{l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})\}}{\partial\boldsymbol{\theta}_G\partial\boldsymbol{\theta}_G^{\mathrm{T}}}$$

and $\hat{\boldsymbol{\theta}}_G$ is the mode of $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})$. The optimal graph is chosen such that the criterion $\mathrm{BNRC}_{\mathrm{dynamic}}$ (8) is minimal.

4. Estimation of gene networks

In this section, we show the concrete strategy for estimating a gene network from cDNA microarray time series gene expression data.

4.1. Nonparametric regression

use the basis function approach for constructing the smooth function $m_{jk}(\cdot)$ described in Section 2. In this paper we use B-splines (see De Boor, 1978) as the basis functions. De Boor's algorithm (see, De Boor, 1978, Chapter 10, p. 130 (3)) is a useful method for computing B-splines of any degree. We use 20 B-splines of degree 3 with equidistance knots (see also, Imoto and Konishi, 2003; Dierckx, 1993; Eiler and Marx, 1996 for the details of B-spline). Fig. 3 shows an example of a B-spline smoothed estimate. We consider a graph $\mathrm{gene}_1 \rightarrow \mathrm{gene}_2$ and estimate a functional structure represented by a thin curve based on our method.

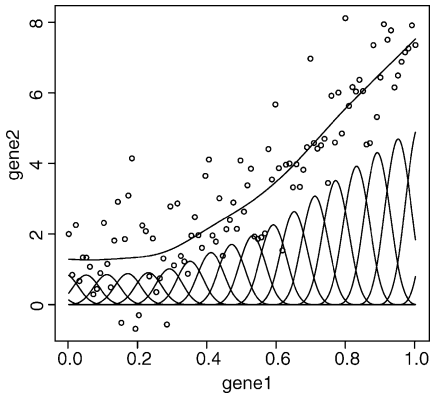


Fig. 3. Fitting a *B*-spline curve (thick curve) to simulated data. Thin curves are weighted *B*-splines. The thick curve is obtained by summing these weighted *B*-splines.

4.2. Prior distribution on the parameter

For the prior distribution on the parameter $\boldsymbol{\theta}_{\mathrm{G}}$, suppose that the parameter vectors $\boldsymbol{\theta}_j$ are independent one another, the prior distribution can then be decomposed as $\pi(\boldsymbol{\theta}_{\mathrm{G}}|\boldsymbol{\lambda}) = \prod_{j=1}^{p} \pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)$. Suppose that the prior distribution $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)$ is factorized as $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) = \prod_{k=1}^{q_j} \pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk})$, where λ_{jk} are hyper parameters. We use a singular M_{jk} variate normal distribution as the prior distribution on $\boldsymbol{\gamma}_{jk}$,

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

where K_{jk} is an $M_{jk} \times M_{jk}$ symmetric positive semidefinite matrix satisfying $\boldsymbol{\gamma}_{jk}^{\mathrm{T}} K_{jk} \boldsymbol{\gamma}_{jk} = \sum_{\alpha=3}^{M_{jk}} (\gamma_{\alpha k}^{(j)} - 2\gamma_{\alpha-1,k}^{(j)} + \gamma_{\alpha-2,k}^{(j)})^2$. This setting of the prior distribution on $\boldsymbol{\theta}_{\mathrm{G}}$ is the same as Imoto et al. (2002a,b) and the details are in those papers.

4.3. Proposed criterion

By using the prior distributions in Section 4.1, the $\mathrm{BNRC}_{\mathrm{dynamic}}$ can be decomposed as follows:

$$\mathrm{BNRC}_{\mathrm{dynamic}} = \sum_{j=1}^{p} \mathrm{BNRC}_{\mathrm{dynamic}}^{(j)}, \tag{9}$$

where $\mathrm{BNRC}_{\mathrm{dynamic}}^{(j)}$ is a local criterion score of jth gene. This decomposition enables us to avoid the complexity of estimating the p-dimensional function. $\mathrm{BNRC}_{\mathrm{dynamic}}^{(j)}$ is defined by

$$\begin{aligned}\mathrm{BNRC}_{\mathrm{dynamic}}^{(j)} = & \\ & -2\log\left\{\int \pi^{(j)} \prod_{i=1}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)d\boldsymbol{\theta}_j\right\} \\ & \approx -2\log\pi^{(j)} - r_j \log\left(\frac{2\pi}{n}\right) \\ & \quad + \log|J_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j)| - 2nl_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}),\end{aligned}$$

where r_j is the dimension of $\boldsymbol{\theta}_j$,

$$l_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}) = \sum_{i=1}^{n} \log g_j(x_{ij}|\boldsymbol{p}_{i-1,j};\boldsymbol{\theta}_j)/n + \log\pi(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)/n,$$

$$J_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j) = -\frac{\partial^2\{l_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X})\}}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^{\mathrm{T}}}$$

and $\hat{\boldsymbol{\theta}}_j$ is the mode of $l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X})$. Here $\pi^{(j)}$ are prior probabilities satisfying

$$\sum_{j=1}^{p} \log\pi^{(j)} = \log\pi(G).$$

We set the prior probability of local structure $\pi^{(j)}$ as

$$\pi^{(j)} = \exp\{-\#(\text{parent genes of } j\text{th gene})\}.$$

4.4. Algorithm for learning network

By using the dynamic Bayesian network and nonparametric regression model together with the proposed criterion, $\mathrm{BNRC}_{\mathrm{dynamic}}$, we can formulate the network learning process as follows: it is clear from (5) and (9) that the optimization of network structure is equivalent to the choices of the parent genes that regulate the target genes. However, it is a time-consuming task when we consider all possible gene combinations as the parent genes. Therefore, we cut down the learning space by selecting candidate parent genes. After this step, a greedy hill-climbing algorithm is employed for finding better networks. Our algorithm can be expressed as follows:

Step 1 (Preprocessing stage). We make the $p \times p$ matrix whose (i, j)th element is the $\mathrm{BNRC}^{(j)}_{\mathrm{dynamic}}$ score of the graph "$\mathrm{gene}_i \rightarrow \mathrm{gene}_j$" and we define the candidate set of parent genes of gene_j that gives small $\mathrm{BNRC}^{(j)}_{\mathrm{dynamic}}$ scores.

Step 2 (Learning stage). For a greedy hill-climbing algorithm, we start form the empty network and repeat the following steps:

Step 2.1: For gene_j, implement one from two procedures that *add* a parent gene, *delete* a parent gene, which gives smaller $\mathrm{BNRC}^{(j)}_{\mathrm{dynamic}}$ score.
Step 2.2: Repeat Step 2.1 until we find the best set of parent genes of jth gene.
Step 2.3: Repeat Step 2.1 and 2.2 for all genes.
Step 2.4: We choose the optimal network that gives the smallest $\mathrm{BNRC}_{\mathrm{dyanmic}}$ score.

5. Computational experiment

5.1. Monte Carlo simulation

Before analyzing real gene expression data, we conduct Monte Carlo simulations to examine the properties of our method. We set an artificial network shown in Fig. 4(a) and suppose functional relationships between nodes in Fig. 4(b). The noise $\varepsilon_{i,j}$ are independently and normally distributed with mean 0 and standard deviation s for $\varepsilon_{i,1}$, $\varepsilon_{i,2}$, $\varepsilon_{i,3}$, $\varepsilon_{i,6}$, $\varepsilon_{i,9}$ and $\varepsilon_{i,10}$, $2s$ for $\varepsilon_{i,4}$, $\varepsilon_{i,5}$ and $\varepsilon_{i,7}$ and $s/2$ for $\varepsilon_{i,8}$, respectively. As for generating the data, we set $x_{1,1} = 0$ as a start point and generate 150 observations for each variable. After generating the data, we remove first 50 observations and use remained 100 observations (i.e. $i = 51, \ldots, 150$) for estimating a network. We set s to 8, 12, 16, 20 and generated 1000 datasets for each case. We observed that when s is set to 12, the data seems to be closest to the real microarray data.

Table 1
Result of Monte Carlo simulations

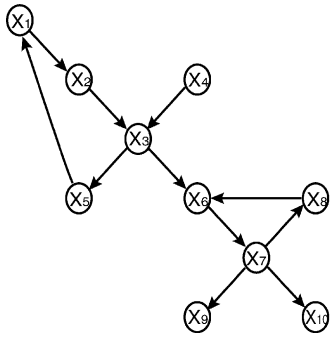
	$s = 8$	$s = 12$	$s = 16$	$s = 20$
(a) Sensitivity	0.996	0.971	0.925	0.877
(b) Specificity	0.697	0.782	0.829	0.851

Table 1 shows "sensitivity" and "specificity" of our method under each setting. Here sensitivity and specificity are defined as follows:

$$\text{sensitivity} = \frac{\#\text{correctly estimated edges}}{\#\text{all edges in the target network}},$$

$$\text{specificity} = \frac{\#\text{correctly estimated edges}}{\#\text{all estimated edges}}.$$

It may seem somewhat strange that the results of the data with large noise indicates high specificity. We presume the reason is that when the volume of system noise is small, most variables seem to be related each other. Therefore, the number of false positives for the data with large noise is relatively smaller than those

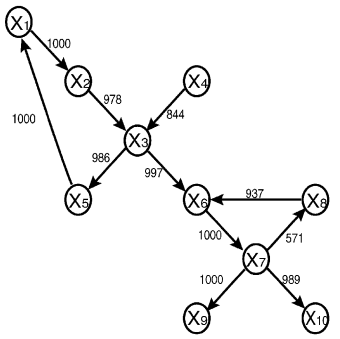


(a)

$$\begin{aligned}
x_{i,1} &= x_{i-1,5} + \varepsilon_{i,1} \\
x_{i,2} &= x_{i-1,1} + \varepsilon_{i,2} \\
x_{i,3} &= -0.6x_{i-1,2} + x_{i-1,4} + \varepsilon_{i,3} \\
x_{i,4} &= \varepsilon_{i,4} \\
x_{i,5} &= -x_{i-1,3} + \varepsilon_{i,5} \\
x_{i,6} &= -x_{i-1,3} + x_{i-1,8} + \varepsilon_{i,6} \\
x_{i,7} &= x_{i-1,6} + \varepsilon_{i,7} \\
x_{i,8} &= -4\sqrt{|x_{i-1,7}|} + 20 + \varepsilon_{i,8} \\
x_{i,9} &= -2x_{i-1,7} + \varepsilon_{i,9} \\
x_{i,10} &= 0.1x^2_{i-1,7} - 50 + \varepsilon_{i,10}
\end{aligned}$$

(b)

Fig. 4. Monte Carlo simulation. (a) Target network, (b) functional structure.



(a)

edge	#
$X_3 \rightarrow X_4$	18
$X_3 \rightarrow X_8$	228
$X_5 \rightarrow X_4$	36
$X_5 \rightarrow X_8$	44
$X_6 \rightarrow X_4$	33
$X_6 \rightarrow X_8$	76
$X_8 \rightarrow X_1$	271
$X_8 \rightarrow X_2$	12
$X_8 \rightarrow X_4$	16
$X_8 \rightarrow X_5$	46
$X_8 \rightarrow X_7$	91
$X_8 \rightarrow X_9$	251

(b)

Fig. 5. Result of the Monte Carlo simulations ($s = 12$). (a) The number of correctly estimated edges, (b) the number of wrongly estimated edges.

for the data with small noise. Note that the resulting network of the data with small noise contains more true positives than that of the data with large noise.

Fig. 5 shows a result of our simulations for $s = 12$. Each number next to an edge indicates how often the edge appeared in the resulting 1000 networks. Fig. 5(b) is the list of false positives. Edges that are estimated less than 10 times are ignored. We succeeded in constructing a cyclic regulation $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_5 \rightarrow X_1$ in 966 networks. Most false positive edges exist around X_8, because the function between X_7 and X_8 is somewhat complex and indifferentiable at $X_7 = 0$. Also, our method assumes the relationship between genes is smooth. However our method estimated this relation 571 times in 1000 networks. We observe that our method works well when the true network contains even cyclic regulations and nonlinear complex dependencies.

5.2. Real data application

We demonstrate our proposed method through the analysis of the *S. cerevisiae* cell cycle gene expression data collected by Spellman et al. (1998). We also apply the Bayesian network and nonparametric regression model (Imoto et al., 2002a,b) and compare the results. This data contains two short time series (two time points; cln3, clb2) and four medium time series (18, 24, 17 and 14 time points; alpha, cdc15, cdc28 and elu). In the estimation of a gene network, we use four medium time series. For combining four time series, we ignore the first observation of the target gene and last one of parent genes for each time series when we fit the nonparametric regression model.

At first, we focus on the cell cycle pathway compiled in KEGG database (http://www.genome.ad.jp/kegg/). The target network is around CDC28 (YBR160w; cyclin-dependent protein kinase). This

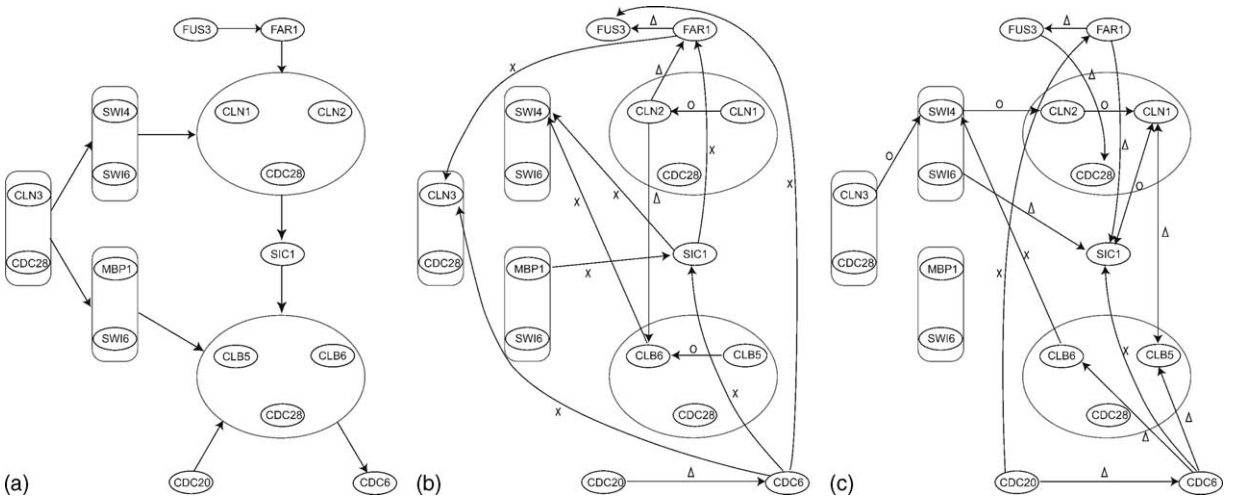


Fig. 6. Cell cycle pathway compiled in KEGG. (a) Target pathway (b) Result of the Bayesian network (c) Result of the proposed method.

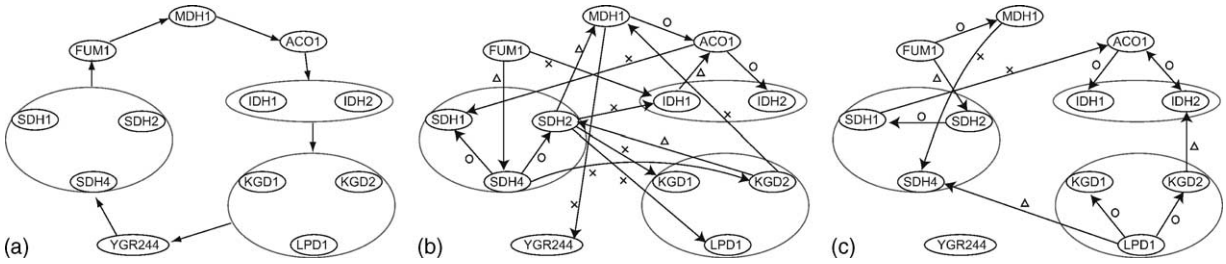


Fig. 7. Metabolic pathway reported by DeRisi et al. (1997). (a) Target pathway (b) Result of the Bayesian network (c) Result of the proposed method.

network contains 45 genes and the partial pathway registered in KEGG is shown in Fig. 6(a). Fig. (b) and (c) are the resulting networks of Imoto et al. (2002a,b) and the proposed method respectively. The edges in the dotted circles can be considered as the correct edges. We can model some correct relations by using the proposed method. We denote the correct estimation by the circle next to edge. The triangle represents either a misdirected edge or an edge skipping at most one gene. The Christ-cross is wrong estimation.

Our second example is the metabolic pathway reported by DeRisi et al. (1997). This network contains 57 genes and the target pathway is partially shown in Fig. 7(a). It is difficult to estimate the metabolic pathway from cDNA microarray data. However, our model can detect some correct relations.

Comparing with the Bayesian network and nonparametric regression, the number of false positives of the proposed method in Figs. 6(c) and 7(c) is much smaller than those in Figs. 6(b) and 7(b). We observed that the Bayesian network and nonparametric regression can work well in many cases. However, when there is a cyclic gene regulation, the Bayesian network and nonparametric regression model tends to estimate many false positives in the cyclic regulation. In such case, the proposed method can reduce the number of false positives and estimate gene regulations effectively.

6. Discussion

In this paper, we proposed a new statistical gene network estimation method based on the dynamic Bayesian network and nonparametric regression model. The advantages of our proposed method compared with other network estimation method such as the Bayesian and Boolean networks are as follows: Our model can take time information into account naturally. Our model can analyze the microarray data as the continuous data without the extra data pretreatments such as discretization. Even nonlinear relations can be detected and modeled by our proposed method.

The simulation of genetic system is one of the central topics in systems biology. Since the simulation is based on the biological knowledge, our network estimation method can support the biological simulation by constructing the unknown regulations. In this paper, we only demonstrate the model based on the first-order Markov relation between time points described in Fig. 1. However, the relationship between time points is arbitrary and we can choose the time dependency structure based on our proposed criterion. Furthermore, when some genes form a protein complex, their expression levels probably change simultaneously. Therefore, the use of a direct graph for representing a protein-protein complex is not suitable. We would like to investigate these topics as our future paper.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (Eds.), Proceedings of the 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

Bilmes, J., 2000. Dynamic bayesian multinets. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Stanford University, CA, pp. 38–45.

Burnham, K., Anderson, D., 1998. Model Selection and Inference, A Practical Information-Theoretical Approach. Springer-Verlag, New York.

Chen, T., He, H., Church, G., 1999. Modeling gene expression with differential equations. Pacif. Symp. Biocomput. 4, 29–40.

Davison, A., 1986. Approximate predictive likelihood. Biometrika 73, 323–332.

De Boor, C., 1978. A Practical Guide to Splines. Springer-Verlag, Berlin.

De Hoon, M., Imoto, S., Kobayashi, K., Ogasawara, N., Miyano, S., 2003. Inferring gene regulatory networks from time-ordered gene expression data of *bacillus subtilis* using differential equations. Pacif. Symp. Biocomput. 8, 17–28.

DeRisi, J., Lyer, V., Brown, P., 1997. Exploring the metabolic and gene control of gene expression on a genomic scale. Science 278, 680–686.

Dierckx, P., 1993. Curve and Surface Fitting with Splines. Oxford.

Eiler, P., Marx, B., 1996. Flexible smoothing with *b*-splines and penalties (with discussion). Stat. Sci. 11, 89–121.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using bayesian network to analyze expression data. J. Comp. Biol. 7, 601–620.

Friedman, N., Murphy, K., Russell, S., 1998. Learning the structure of dynamic probabilistic networks. In: proceedings of the Conference on Uncertainty in Artificial Intelligence, University of Wisconsin Business School, Madison, WI, pp. 139–147.

Imoto, S., Goto, T., Miyano, S., 2002a. Estimation of genetic networks and functional structures between genes by using bayesian network and nonparametric regression. Pacif. Symp. Biocomput. 7, 175–186.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., Miyano, S., 2002b. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. J. Bioinformat. Computat. Biol. 1, 231–252.

Imoto, S., Konishi, S., 2003. Selection of smoothing parameters in b-spline nonparametric regression models using information criteria. Ann. Inst. Stat. Mathemat. 55, 671–687.

Konishi, S., 1999. Statistical model evaluation and information criteria. In: Ghosh, S. (Ed.), Multivaliate Analysis, Design of Experiments and Survey Sampling. Marcel Dekker, New York, pp. 369–399.

Konishi, S., Ando, T., Imoto, S., 2004. Bayesian information criteria and smoothing parameter selection in radial basis function networks. Biometrika 91, 27–43.

Konishi, S., Kitagawa, G., 1996. Generalized information criteria in model selection. Biometrika 83, 875–890.

Ong, I.M., Glasner, J.D., Page, D., 2002. Modelling regulatory pathways in *e. coli* from time series expression profiles. Bioinformatics 18, S241–S248.

Pe'er, D., Regev, A., Elidan, G., Friedman, N., 2001. Inferring subnetworks from perturbed expression profiles. Bioinformatics 17, S215–S224.

Someren, E., Wessels, L., Reinders, M., 2002. Linear modeling of genetic networks from experimental data. Bioinformatics 18, S355–S366.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Tinerey, L., Kadane, J., 1986. Accurate approximations for posterior moments and marginal densities. J. Am. Statist. Assoc. 81, 82–86.