Bayesian Network and Nonparametric Heteroscedastic Regression for Nonlinear Modeling of Genetic Network

Seiya Imoto[1], Kim Sunyong[1], Takao Goto[1], Sachiyo Aburatani , Kousuke Tashiro ,
Satoru Kuhara and Satoru Miyano[1]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan
{imoto, sunk, takao, miyano}@ims.u-tokyo.ac.jp

Graduate School of Genetic Resources Technology, Kyushu University
6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan
{sachiyo, ktashiro, kuhara}@grt.kyushu-u.ac.jp

Abstract

We propose a new statistical method for constructing a genetic network from microarray gene expression data by using a Bayesian network. An essential point of Bayesian network construction is in the estimation of the conditional distribution of each random variable. We consider fitting nonparametric regression models with heterogeneous error variances to the microarray gene expression data to capture the nonlinear structures between genes. A problem still remains to be solved in selecting an optimal graph, which gives the best representation of the system among genes. We theoretically derive a new graph selection criterion from Bayes approach in general situations. The proposed method includes previous methods based on Bayesian networks. We demonstrate the effectiveness of the proposed method through the analysis of Saccharomyces cerevisiae gene expression data newly obtained by disrupting 100 genes.

1. Introduction

Due to the development of the microarray technology, constructing genetic network receives a large amount of attention in the fields of molecular biology and bioinformatics [3, 4, 5, 14, 15, 17, 22, 28]. However, the dimensionality and complexity of the data disturb the progress of the microarray gene expression data analysis. That is to say, the information that we want is buried in a huge amount of the data with noise. In this paper, we propose a new statistical method for constructing a genetic network that can capture even the nonlinear relationships between genes clearer.

A Bayesian network [7, 23] is an effective method in modeling phenomena through the joint distribution of a large number of random variables. In recent years, some interesting works have been established in constructing genetic networks from microarray gene expression data by using Bayesian networks. Friedman and Goldszmidt [12, 13, 14] discretized the expression values and assumed multinomial distributions as the candidate statistical models. Pe'er *et al.* [28] investigated the threshold value for discretizing. On the other hand, Friedman *et al.* [15] pointed out that the discretizing probably loses information of the data. In fact, the number of discretizing values and the thresholds are unknown parameters, which have to be estimated from the data. The resulted network strongly depends on their values. Then Friedman *et al.* [15] considered fitting linear regression models, which analyze the data in the continuous (see also [20]). However, the assumption that the parent genes depend linearly on the objective gene is not always guaranteed. Imoto *et al.* [22] proposed the use of nonparametric additive regression models (see also [16, 18]) for capturing not only linear dependencies but also nonlinear structures between genes. In this paper, we propose a method for constructing the genetic network by using Bayesian networks and the nonparametric heteroscedastic regression, which is more resistant to the effect of outliers.

Once we set the graph, we have to evaluate its goodness or closeness to the true graph, which is completely unknown. Hence, the construction of a suitable criterion becomes the center of attention of statistical genetic network modeling. Friedman and Goldszmidt [14] used the BDe criterion, which was originally derived by [21] for choos-

ing a graph. The BDe criterion only evaluates the Bayesian network based on the multinomial distribution model and Dirichlet priors. However, Friedman and Goldszmidt [14] kept the unknown hyper parameters in Dirichlet priors and we only set up the values experimentally. We investigate the graph selection problem as a statistical model selection or evaluation problem and theoretically derive a new criterion for choosing a graph using the Bayes approach (see [6]). The proposed criterion automatically optimizes all parameters in the model and gives the optimal graph. In addition, our proposed method includes the previous methods for constructing genetic network based on Bayesian network. To show the effectiveness of the proposed method, we analyze gene expression data of *Saccharomyces cerevisiae* newly obtained by disrupting 100 genes.

2. Bayesian Network and Nonparametric Heteroscedastic Regression Model

2.1. Nonlinear Bayesian network model

Suppose that we have n sets of array data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of p genes, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $\boldsymbol{x}'$ denotes the transpose of $\boldsymbol{x}$. In the Bayesian network framework, we consider a directed acyclic graph and Markov assumption between nodes. The joint density function is then decomposed into the conditional density of each variable, that is,

$$f(x_{i1}, \ldots, x_{ip}) = \prod_{j=1}^{p} f_j(x_{ij} | \boldsymbol{p}_{ij}) \tag{1}$$

where $\boldsymbol{p}_{ij} = (p_{i1}^{(j)}, \ldots, p_{iq_j}^{(j)})'$ are q_j-dimensional parent observation vectors of x_{ij} in the graph. When gene$_2$ and gene$_3$ are parent genes of gene$_1$, we see $\boldsymbol{p}_{i1} = (x_{i2}, x_{i3})'$, $(i = 1, \ldots, n)$. Through formula (1), the focus of interest in statistical modeling by Bayesian networks is how we can construct the conditional densities, f_j. We assume that the conditional densities, f_j, are parameterized by the parameter vectors $\boldsymbol{\theta}_j$, and the information is extracted from these probabilistic models.

Imoto *et al.* [22] proposed the use of nonparametric regression strategy for capturing the nonlinear relationships between x_{ij} and $\boldsymbol{p}_{ij}$ and suggested that there are many nonlinear relationships between genes and the linear model hardly achieves a sufficient result. In many cases, this method can capture the objective relationships very well. When the data, however, contain outliers especially near the boundary of the domain $\{\boldsymbol{p}_{ij}\}$, nonparametric regression models sometimes lead to unsuitable smoothed estimates, i.e., the estimated curve exhibits some spurious waviness due to the effects of the outliers. Since what is estimated

is the system of a living nature, a too complicated relationship is unsuitable. In fact, this inappropriate case unfortunately sometimes occurs in the analysis of real data. To avoid this problem, we consider fitting a nonparametric regression model with heterogeneous error variances

$$x_{ij} = m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)}) + \varepsilon_{ij} \tag{2}$$

where ε_{ij} depends independently and normally on mean 0 and variance σ_{ij}^2 and $m_{jk}(\cdot)$ is a smooth function from $\mathbb{R}$ to $\mathbb{R}$. Here $\mathbb{R}$ denotes a set of real numbers. This model includes Imoto *et al.* [22]'s model and, clearly, the linear regression model as special cases. In general, each smooth function $m_{jk}(\cdot)$ is characterized by the n values $m_{jk}(p_{1k}^{(j)}), \ldots, m_{jk}(p_{nk}^{(j)})$ and the system (2) contains $(n \times q_j + n)$ parameters. Then the number of the parameters in the model is much larger than the number of observations and it has a tendency toward unstable parameter estimates. In this paper, we construct the smooth function $m_{jk}(\cdot)$ by the basis functions approach

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

where $\gamma_{1k}^{(j)}, \ldots, \gamma_{M_{jk}k}^{(j)}$ are unknown coefficient parameters and $b_{1k}^{(j)}(\cdot), \ldots, b_{M_{jk}k}^{(j)}(\cdot)$ are basis functions. From this representation, the parameters $m_{jk}(p_{1k}^{(j)}), \ldots, m_{jk}(p_{nk}^{(j)})$ are reparameterized by the M_{jk} coefficient parameters $\gamma_{1k}^{(j)}, \ldots, \gamma_{M_{jk}k}^{(j)}$.

We strongly recommend the use of nonparametric regression instead of linear regression, because linear regression cannot decide the direction of the Bayes causality or leads to the wrong direction in many cases. We show the advantage of the proposed model compared with linear regression through a simple example. Suppose that we have data of gene$_1$ and gene$_2$ in Figure 1 (a). We consider the two models gene$_1 \rightarrow$ gene$_2$ and gene$_2 \rightarrow$ gene$_1$, and obtain the smoothed estimates shown in Figure 1 (b) and (c), respectively. We decide that the model (b: gene$_1 \rightarrow$ gene$_2$) is better that (c: gene$_2 \rightarrow$ gene$_1$) by the proposed criterion, which is derived in a later section (the scores of the models are (b) 120.6 (c) 134.8). Since we generated this data from the true graph gene$_1 \rightarrow$ gene$_2$, our method yields the correct result. However, if we fit the linear regression model to this data, the model (c) is chosen (the scores are (b) 156.0 (c) 135.8). The method, which is based on linear regression, yields an incorrect result in this case.

Consider the case that the relationship is almost linear. Our method and linear regression can fit the data appropriately. However, it is clearly difficult to decide the direction of Bayes causality. In such a case, the direction is not strict.
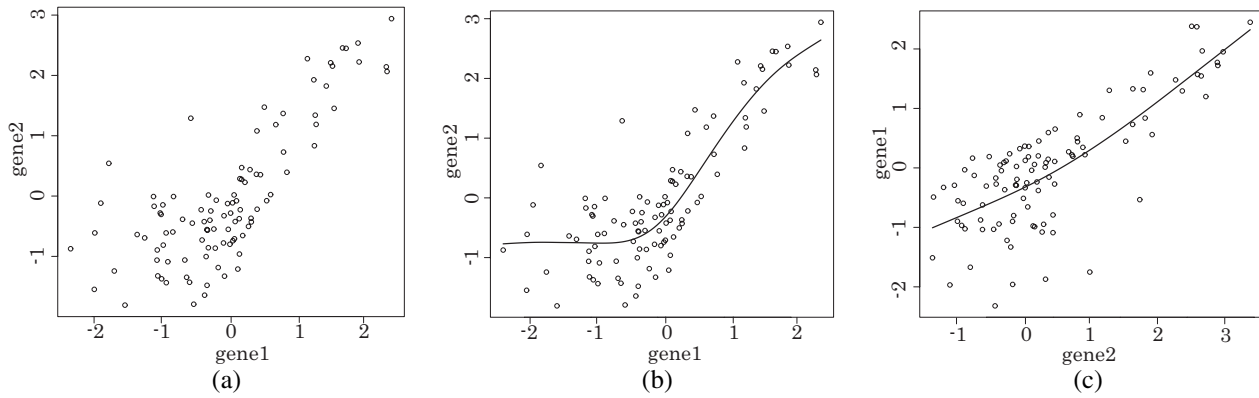
Figure 1. Simulated data: The true causality is gene$_1$ $\rightarrow$ gene$_2$. (a) Scatter plot of the simulated data. (b) Smoothed curve of the graph gene$_1$ $\rightarrow$ gene$_2$. (c) Smoothed curve of the graph gene$_2$ $\rightarrow$ gene$_1$. These curves are obtained by the proposed method.

In the error variances, σ_{ij}^2, we assume the structures,

$$\sigma_{ij}^2 = \kappa_{ij}^{-1}\sigma_j^2, \qquad \kappa = 1 \qquad ; \qquad = 1 \qquad (3)$$

where $\kappa_{1j} \cdots \kappa_{nj}$ are constants and σ_j^2 is an unknown parameter. By setting up the constants $\kappa_{1j} \cdots \kappa_{nj}$ in reflecting the feature of the error variances, we can represent the heteroscedasticity of the data. Combining (2) and (3), we obtain a nonparametric regression model with heterogeneous error variances

$$f_j(x_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\gamma}_j,\sigma_j^2) = \frac{\kappa_{ij}^{1/2}}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{\kappa_{ij}}{2\sigma_j^2}\{x_{ij} - \sum_{k=1}^{q_j}\boldsymbol{\gamma}_{jk}\boldsymbol{b}_{jk}(p_{ik}^{(j)})\}^2\right] \qquad (4)$$

where $\boldsymbol{\gamma}_{jk}$ and $\boldsymbol{b}_{jk}(p_{ik}^{(j)})$ are M_{jk}-dimensional vectors given by, respectively, $\boldsymbol{\gamma}_{jk} = (\gamma_{1k}^{(j)} \cdots \gamma_{M_{jk}k}^{(j)})$ and $\boldsymbol{b}_{jk}(p_{ik}^{(j)}) = (b_{1k}^{(j)}(p_{ik}^{(j)}) \cdots b_{M_{jk}k}^{(j)}(p_{ik}^{(j)}))$. If the j-th gene has no parent genes in the graph, we specify the model based on the normal distribution with mean μ_j and variance σ_j^2. Hence, we define the nonlinear Bayesian network model

$$f(\boldsymbol{x}_i;\boldsymbol{\theta}_G) = \prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\theta}_j) \qquad (5)$$

where $\boldsymbol{\theta}_G = (\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_p)$ is the parameter vector included in the graph G and $\boldsymbol{\theta}_j$ is the parameter vector in the conditional density f_j, that is, we see $\boldsymbol{\theta}_j = (\boldsymbol{\gamma}_j, \sigma_j^2)$ or $\boldsymbol{\theta}_j = (\mu_j, \sigma_j^2)$.

2.2. Criterion for choosing graph

Once we set a graph, the statistical model (5) based on the Bayesian network and nonparametric regression can be constructed and be estimated by a suitable procedure. However, the problem that still remains to be solved is how we can choose the optimal graph, which gives a best approximation of the system underlying the data. Notice that we cannot use the likelihood function as a model selection criterion, because the value of likelihood becomes large in a more complicated model. Hence, we need to consider the statistical approach based on the generalized or predictive error, Kullback-Leibler information, Bayes approach and so on (see e.g., [1, 24, 25] for the statistical model selection problem). In this section, we construct a criterion for evaluating a graph based on our model (5) from Bayes approach.

The posterior probability of the graph is obtained by the product of the prior probability of the graph, π_G, and the marginal probability of the data. By removing the standardizing constant, the posterior probability of the graph is proportional to

$$\pi(\ |\boldsymbol{X}_n) = \pi_G \int \prod_{i=1}^{n} f(\boldsymbol{x}_i;\boldsymbol{\theta}_G) \ \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) \ d\boldsymbol{\theta}_G \qquad (6)$$

where $\boldsymbol{X}_n = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)$ is an $n \times p$ gene profile matrix, $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ is the prior distribution on the parameter $\boldsymbol{\theta}_G$ satisfying $\log \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(\ )$ and $\boldsymbol{\lambda}$ is the hyper parameter vector. Under Bayes approach, we can choose the optimal graph such that $\pi(\ |\boldsymbol{X}_n)$ is maximum. A crucial problem for constructing a criterion based on the posterior probability of the graph is the computation of the high dimensional integration (6). Heckerman and Geiger [20] used the conjugate priors for solving the integral and gave a closed-form

solution. To compute this high dimensional integration, we use Laplace's approximation [9, 19, 31] for integrals

$$\int \prod_{i=1}^{n} (\boldsymbol{x}_i; \boldsymbol{\theta}_G) \, (\boldsymbol{\theta}_G|\boldsymbol{\lambda}) \, \boldsymbol{\theta}_G$$

$$= \frac{(2\pi)^{r/2}}{|\, (\hat{\boldsymbol{\theta}}_G)|^{1/2}} \exp\{ (\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}_n)\}\{1 + \, _p(\, ^{-1})\}$$

where $\quad$ is the dimension of $\boldsymbol{\theta}_G$, $\quad (\boldsymbol{\theta}_G|\boldsymbol{X}_n) = \sum_{i=1}^{n} \log \, (\boldsymbol{x}_i; \boldsymbol{\theta}_G) \, + \log \, (\boldsymbol{\theta}_G | \, \boldsymbol{\lambda}) \quad (\boldsymbol{\theta}_G) = -\, ^2\{ \, (\boldsymbol{\theta}_G|\boldsymbol{X}_n)\} \, \boldsymbol{\theta}_G \, \boldsymbol{\theta}_G$ and $\hat{\boldsymbol{\theta}}_G$ is the mode of $(\boldsymbol{\theta}_G|\boldsymbol{X}_n)$. Then we define the Bayesian network and nonparametric heteroscedastic regression criterion, named BNRC$_{eter}$, for selecting a graph

$$\mathrm{BNRC}_{eter}(\quad)$$

$$= -2\log \quad_G \int \prod_{i=1}^{n} (\boldsymbol{x}_i; \boldsymbol{\theta}_G) \, (\boldsymbol{\theta}_G|\boldsymbol{\lambda}) \, \boldsymbol{\theta}_G$$

$$-2\log \quad_G - \quad \log(2\pi) + \log | \quad (\hat{\boldsymbol{\theta}}_G)|$$

$$-2 \quad (\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}_n) \tag{7}$$

The optimal graph is chosen such that the criterion BNRC$_{eter}$ (7) is minimal. The merit of the use of the Laplace method is that it is not necessary to consider the use of the conjugate prior distribution. Hence the modeling in the larger classes of distributions of the model and prior is attained.

Suppose that the parameter vectors $\boldsymbol{\theta}_j$ are independent one another, the prior distribution can be decomposed into $(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \prod_{j=1}^{p} \, _j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)$. Therefore, $\log | \quad (\boldsymbol{\theta}_G|\boldsymbol{X}_n)|$ and $(\boldsymbol{\theta}_G|\boldsymbol{X}_n)$ in (7) result in, respectively,

$$\log | \quad (\boldsymbol{\theta}_G|\boldsymbol{X}_n)| = \sum_{j=1}^{p} \log \quad - \frac{\, ^2 \, _j(\boldsymbol{\theta}_j|\boldsymbol{X}_n)}{\boldsymbol{\theta}_j \, \boldsymbol{\theta}_j}$$

$$(\boldsymbol{\theta}_G|\boldsymbol{X}_n) = \sum_{j=1}^{p} \, _j(\boldsymbol{\theta}_j|\boldsymbol{X}_n)$$

where $\quad _j(\boldsymbol{\theta}_j|\boldsymbol{X}_n) = \quad \log \, _j(\, _{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\theta}_j) \quad + \log \, _j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) \quad$. Here $\boldsymbol{\lambda}_j$ is the hyper parameter vector. Hence by defining

$$\mathrm{BNRC}^{(j)}_{eter}$$

$$= -2\log \quad \int \, _{L_j} \prod_{i=1}^{n} \, _j(\, _{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\theta}_j) \, _j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) \, \boldsymbol{\theta}_j$$

where $\quad _{L_j}$ are prior probabilities satisfying $\sum_{j=1}^{p} \log \, _{L_j} = \log \, _G$, the BNRC$_{eter}$ score is given by the sum of the local scores

$$\mathrm{BNRC}_{eter} = \sum_{j=1}^{p} \mathrm{BNRC}^{(j)}_{eter} \tag{8}$$
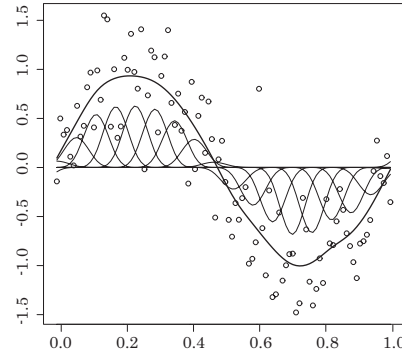

Figure 2. The fitted curve to simulated data: The thin curves are B-splines that are weighted by coefficients and the thick curve is the smoothed estimate that is otained by the linear combination of weighted B-splines.

The smoothed estimates based on nonparametric heteroscedastic regression are obtained by replacing the parameters $\boldsymbol{\gamma}_j$ by $\hat{\boldsymbol{\gamma}}_j$. Noticed that we derive the criterion, BNRC$_{eter}$, under the assumption $\log \, (\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \, (\quad)$. If we use the prior density satisfying $\log \, (\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \, (1)$, the BNRC$_{eter}$ score results in Schwarz's criterion known as BIC or SIC [30]. In such case, the mode $\hat{\boldsymbol{\theta}}_G$ is equivalent to the maximum likelihood estimate.

3. Estimating Genetic Network

3.1. Nonparametric regression

In this section we present the method for constructing genetic network in practice based on the proposed method described above. First we would like to mention the nonparametric regression model. In the additive model, we construct each smooth function $\quad _{jk}(\cdot)$ by B-splines [10, 22]. Figure 2 is an example of B-splines smoothed curve. The thin curves are B-splines that are weighted by coefficients and thick line is a smoothed curve that is obtained by the linear combination of weighted B-splines.

In the error variances, we consider the heteroscedastic regression model and assume the structure (3). Choosing constants $\quad _{1j} \quad _{nj}$ is an important problem for capturing the heteroscedasticity of the data. In this paper, we set the weights

$$_{ij} = \, (\boldsymbol{p}_{ij}; \, _j) = \exp\{- \, _j ||\boldsymbol{p}_{ij} - \boldsymbol{p}_j||^2 \, 2 \, _j^2\} \tag{9}$$

where $\quad _j$ is a hyper parameter, $\boldsymbol{p}_j = \sum_{i=1}^{n} \boldsymbol{p}_{ij}$ and $\, _j^2 = \sum_{i=1}^{n} ||\boldsymbol{p}_{ij} - \boldsymbol{p}_j||^2 \quad _j$. If we set $\, _j = 0$, the weights

are $\beta_{1j} = \cdots = \beta_{nj} = 1$ and the model has homogeneous error variances. If we use a large value of β_j, the error variances of the data, which exist near the boundary on the domain of the parent variables, are large. Hence, if there are outliers near the boundary, we can reduce their effect and gain the suitable smoothed estimates by using the appropriate value of β_j.

3.2. Priors

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 $\boldsymbol{\gamma}_{jk}$,

$$\pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk}) = \left(\frac{\lambda_{jk}}{2\pi}\right)^{-(M_{jk}-2)/2} |K_{jk}|_+^{1/2}$$
$$\times \exp\left(-\frac{\lambda_{jk}}{2}\boldsymbol{\gamma}_{jk}' K_{jk}\boldsymbol{\gamma}_{jk}\right) \quad (10)$$

where K_{jk} is an $M_{jk} \times M_{jk}$ symmetric positive semidefinite matrix satisfying $\boldsymbol{\gamma}_{jk}' K_{jk}\boldsymbol{\gamma}_{jk} = \sum_{\ell=1}^{M_{jk}}(\gamma_{\ell,k}^{(j)} - 2\gamma_{\ell-1,k}^{(j)} + \gamma_{\ell-2,k}^{(j)})^2$.

Next we consider the prior probability of the graph π_G. Friedman and Goldszmit [14] employed the prior based on the MDL encoding of the graph. In our context, the marginal probability of the data is equivalent to the type II likelihood adjusted by the hyper parameters. Thus we set the prior probability of the graph, π_G,

$$\pi_G = \exp\{-(\text{No of hyper parameters})\}$$
$$= \prod_{j=1}^{p} \exp\{-(q_j+1)\} = \prod_{j=1}^{p} L_j$$

The justification of this prior is based on Akaike's Bayesian information criterion, known as ABIC [2], and Akaike's information criterion, AIC [1].

3.3. Criterion

We derived the criterion, BNRC$_{eter}$, for choosing the graph in a general framework. By using the equation (8), the BNRC$_{eter}$ score of the graph can be obtained by the sum of the local scores, BNRC$_{eter}^{(j)}$. The result is summarized in the following theorem.

Theorem 1. Let $(\boldsymbol{x}_i; \boldsymbol{\theta}_G)$ be a Bayesian network and nonparametric heteroscedastic regression model given by (5), and let $\pi(\boldsymbol{\gamma}_{jk}|\lambda_{jk})$ be the prior densities on the parameters $\boldsymbol{\gamma}_{jk}$ defined by (10). Then a criterion for evaluating graph is given by BNRC$_{eter} = \sum_{j=1}^{p}$ BNRC$_{eter}^{(j)}$, where

$$\text{BNRC}_{eter}^{(j)} = 2(q_j+1) - (\sum_{k=1}^{q_j} M_{jk}+1)\log(2\pi)$$

$$- \sum_{i=1}^{n}\log\sigma_{ij} + n\log(2\pi\hat{\sigma}_j^2) +$$

$$+ \sum_{k=1}^{q_j}\{\log|K_{jk}| - M_{jk}\log(\lambda\hat{\sigma}_j^2)\} - \log(2\pi\hat{\sigma}_j^2)$$

$$+ \sum_{k=1}^{q_j}\{(M_{jk}-2)\log(2\pi\hat{\sigma}_j^2\lambda_{jk}) - \log|K_{jk}|_+$$

$$+ \lambda_{jk}\hat{\boldsymbol{\gamma}}_{jk}' K_{jk}\hat{\boldsymbol{\gamma}}_{jk}/\hat{\sigma}_j^2\}$$

with

$$H_{jk} = B_{jk}' \Sigma_j B_{jk} + \lambda_{jk} K_{jk}; \quad (M_{jk} \times M_{jk})$$
$$B_{jk} = (\boldsymbol{b}_{1k}^{(j)}(\xi_{1k}^{(j)}) \cdots \boldsymbol{b}_{M_{jk}k}^{(j)}(\xi_{nk}^{(j)}))'; \quad (n \times M_{jk})$$
$$\Sigma_j = \text{diag}(\sigma_{1j} \cdots \sigma_{nj}); \quad (n \times n)$$
$$\hat{\sigma}_j^2 = \sum_{i=1}^{n}\sigma_{ij}\{y_{ij} - \sum_{k=1}^{q_j}\hat{\boldsymbol{\gamma}}_{jk}'\boldsymbol{b}_{jk}(\xi_{ik}^{(j)})\}^2$$

Here we approximate the Hessian matrix by

$$\log\left|-\frac{\partial^2\log\pi_j(\boldsymbol{\theta}_j|\boldsymbol{X}_n)}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j'}\right| \approx \sum_{k=1}^{q_j}\log\left|-\frac{\partial^2\log\pi_j(\boldsymbol{\theta}_j|\boldsymbol{X}_n)}{\partial\boldsymbol{\gamma}_{jk}\partial\boldsymbol{\gamma}_{jk}'}\right|$$
$$+ \log\left|-\frac{\partial^2\log\pi_j(\boldsymbol{\theta}_j|\boldsymbol{X}_n)}{\partial(\sigma_j^2)^2}\right|$$

□

3.4. Learning network

In the Bayesian network literature, it is shown that determining the optimal network is an NP-hard problem. In this paper, we use the greedy hill-climbing algorithm for learning network as follows:

Step1: Make the score matrix whose $(\;)$-th element is the BNRC$_{eter}^{(j)}$ score of the graph gene$_i \rightarrow$ gene$_j$.

Step2: For each gene, implement one of three procedures for an edge: "add", "remove", "reverse", which gives the smallest BNRC$_{eter}$.

Step3: Repeat Step2 until the BNRC$_{eter}$ does not reduce.

Generally, the greedy hill-climbing algorithm has many local minima and the result depends on the computational order of variables. To avoid this problem, we permute the computational order of genes and make many candidate learning orders in Step3. Another problem of the learning network is that the search space of the parent genes is enormously wide, when the number of genes is large. Then we restrict the set of the candidate parent genes based on the score matrix, which is given by Step1.

Imoto *et al.* [22] used this learning strategy for learning genetic network and showed the effectiveness of their method by the Monte Carlo simulation method. We also check the efficiencies of our new model through the same Monte Carlo simulations and find improvements due to the nonparametric heteroscedastic regression model, which is newly introduced. We show the effectiveness of the heteroscedastic regression model in the next subsection.

3.5. Hyper parameters

Consider the nonparametric regression model defined in (4). The estimate $\hat{\boldsymbol{\theta}}_j$ is a mode of $\pi_j(\boldsymbol{\theta}_j|\boldsymbol{X}_n)$ and depends on the hyper parameters. In fact, the hyper parameter plays an essential role for estimating the smoothed curve.

In our model, we construct the nonparametric regression model by 20 *B*-splines. We confirmed that the differences of the smoothed estimates against the various number of the basis functions cannot be found visually. Because when we use a somewhat large number of the basis functions, the hyper parameters control the smoothness of the fitted curves. Figure 3 (a1) shows the scatter plot of YGL237C and YEL071W with smoothed estimates for 3 different values of the hyper parameters. The details of the data are shown in later section. Clearly, the smoothed estimate strongly depends on the values of the hyper parameters. Figure 3 (a2) is the behavior of the BNRC$_{eter}$ criterion of the two genes in Figure 3 (a1). We can choose the optimal value of the hyper parameter as the minimizer of the BNRC$_{eter}$ and the optimal smoothed estimate (solid curve in Figure 3 (a1)) can capture the structure between these genes well. The dashed and dotted curves are near the maximum likelihood estimate and the parametric linear fit, respectively.

The effect of the weight constants $\lambda_{1j} \cdots \lambda_{nj}$ is shown in Figure 3 (b1) and (c1). If we use the nonparametric homoscedastic regression model [22], we obtain the dashed curve, which exhibits some spurious waviness due to the effect of the data in the upper-left corner (b1). By adjusting the hyper parameter γ_j in (9), the estimated curve results in the solid curve. The optimal value of γ_j is also chosen by minimizing the BNRC$_{eter}$ criterion (see Figure 3 (b2) and (c2)). Of course, when the smoothed estimate is properly obtained, the optimal value of γ_j tends to zero.

Finally, we show the algorithm for estimating the smoothed curve and optimizing the hyper parameters.
Step1: Fix the hyper parameter λ_j.
Step2: Initialize: $\boldsymbol{\gamma}_{jk} = \boldsymbol{0}$, $k = 1 \cdots p_j$.
Step3: Find the optimal γ_{jk} by repeating Step3-1 and Step3-2
Step3-1: Compute:

$$\boldsymbol{\gamma}_{jk} = (\lambda_{jk} B_{jk} B_{jk} + \mu_{jk} K_{jk})^{-1} \lambda_{jk} B_{jk}$$

$$\times (\boldsymbol{x}_{(j)} - \sum_{k'=k} B_{jk'} \boldsymbol{\gamma}_{jk'})$$

for fixed μ_{jk}.
Step3-2: Evaluate: Repeat Step3-1 against the candidate value of μ_{jk}, and choose the optimal value of μ_{jk}, which minimizes the BNRC$_{eter}^{(j)}$.
Step4: Convergence: Repeat Step3 for $k = 1 \cdots j 1 \cdots j 1$ until a suitable convergence criterion is satisfied.
Step5: Repeat Step1 to Step4 against the candidate value of λ_j, and choose the optimal value of λ_j, which minimizes the BNRC$_{eter}^{(j)}$.

4. Real Data Analysis

In this section we show the effectiveness of our proposed method through the analysis of *Saccharomyces cerevisiae* gene expression data, which is newly obtained by disrupting 100 genes. Our research group has installed a systematic experimental method, which observes changes in the expression levels of genes on a microarray by gene disruption. By using this method, we have launched a project whose purpose is to reveal the gene regulatory networks between the 5871 genes of *Saccharomyces cerevisiae*. Many laboratories have also reported similar projects. We have already collected a large number of expression profiles from gene disruption experiments to evaluate genetic regulatory networks. Over 400 mutants are stocked and gene expression profiles are accumulating.

We monitored the transcriptional level of 5871 genes spotted on a microarray by a scanner. The expression profiles of over 400 disruptants were stored in our database. The standard deviation (SD) of the levels of all genes on a microarray was evaluated. The value of SD represents roughly the experimental error. In our data, we estimated the value of 0.5 as the critical point of the accuracy of experiments. We have evaluated the accuracy of those profiles on the base of the standard deviation of the expression ratio of all genes. 107 disruptants including 68 mutants where the transcription factors were disrupted could be selected from 400 profiles.

We used 100 microarrays and constructed a genetic network of 521 genes from the above data. The 94 transcription factors whose regulating genes have been clearly identified were found. The profiles of the 521 genes in control by those 94 factors were selected from 5871 profiles.

Bas1p and Bas2p also activate expression of three genes in the histidine biosynthesis pathway. In a gcn4 gackground, mutations that abolish the *BAS1* or *BAS2* function lead to a histidine auxotrophy. Previous investigation indicated that Bas1p and Bas2p are DNA binding proteins required for transcription of *HIS4* and these *ADE* genes like
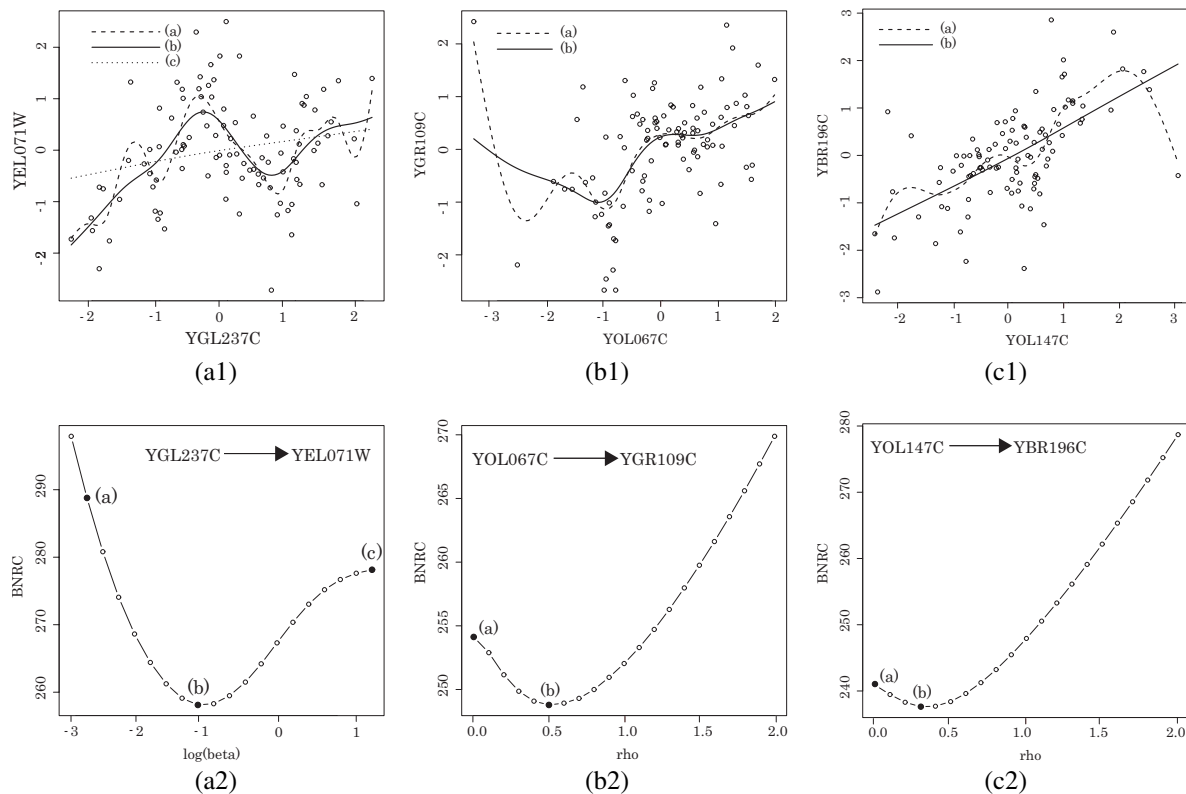
Figure 3. The smoothed estimates by the various values of the hyper parameters. (a1): The effect of hyperparameter β_{jk} in the prior distribution of the coefficients of B-splines. This parameter can control the smoothness of the fited curve. (b1) and (c1): The effect of hyperparamter ρ_j in the parameter of the error variances. This parameter can capture the heteroscedastisity of the data and can reduce the effects of outliers.

GCN4 [8, 11, 29]. In this paper, we made clear that both genetic relation. Figure 4 indicates that those *ADE* genes and histidine biosynethesis genes are related with *BAS1* more directly than *GCN4*. The ribose component of purine ribonucleotides is derived from ribose 5-P, an inter mediate of the pentose phosphate cycle. The atoms of the base moiety are contributed by many compounds. They are added step wise to the preformed ribose. There exist striking interrelationships with the pathway for histidine synthesis.

Studies on the regulation of the purine biosiynthesis pathway in *Saccharomyces cerevisiae* revealed that all the genes encoding enzymes required for AMP de novo biosynthesis are repressed at transcriptional level by the presence of extracellular purines. *ADE* genes are transcriptionally activated as well as some histidine biosynthesis genes. Especially the fact that expression of *HIS4* is related with *ADE* genes were known. In our regulated network, *HIS4* were related with some *ADE* genes closely, and some *HIS* genes are related with *ADE* genes like *HIS4*. The biosynthesis of

the essential amino acid histidine shows in *Saccharomyces cerevisiae* shows close connection to purine metabolism, and our result satisfied this fact.

5. Conclusion

In this paper we proposed a new statistical method for estimating a genetic network from microarray gene expression data by using a Bayesian network and nonparametric regression. The key idea of our method is the use of nonparametric heteroscedastic regression models for capturing nonlinear relationships between genes and heteroscedasticity of the expression data. If we have a network that represents the causal relationship among genes, we can simulate the genetic system on the computer, e.g., Genomic Object Net [26, 27]. In this stage, it is required that the relationships between genes are suitably estimated. In this sense, the proposed heteroscedastic model can give an essential
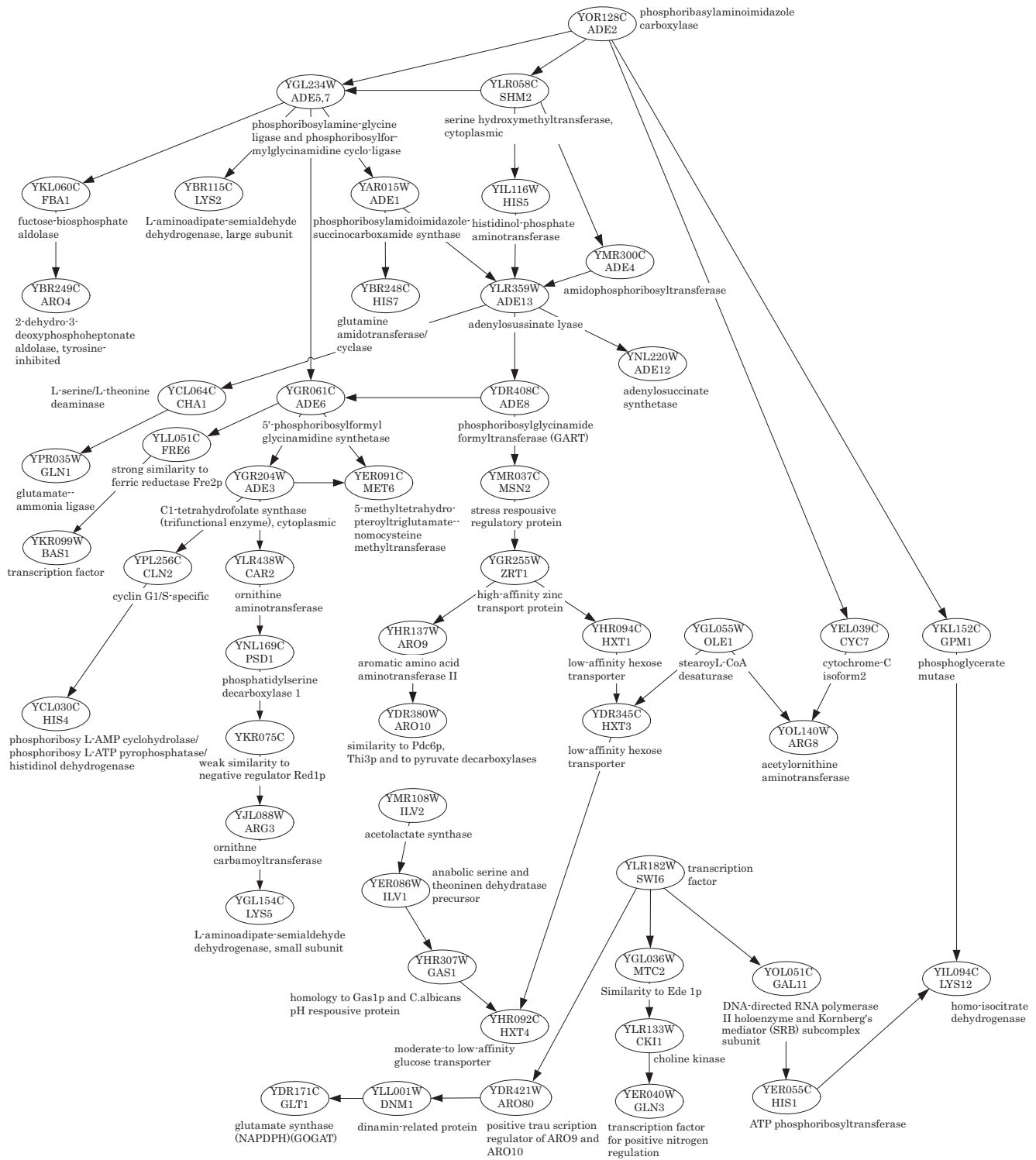
Figure 4. The resulting partial network of the analysis of 521 *Saccharomyces cerevisiae* **genes.**

improvement, because the previous models sometimes lead to unsuitable estimates of the systems. We consider the simulation of biological system as a future work.

An essential problem for network construction is the evaluation of the graph. We investigated this problem as a statistical model selection or evaluation problem and derived the new criterion for selecting graph from Bayes approach. Our method covers the previous methods for constructing genetic networks by using Bayesian networks and improves them in the theoretical and methodological senses. The proposed method successfully extracts the effective information and we can find these information in the resulting genetic network visually. We use the simple greedy algorithm for learning network. However, this algorithm needs much time for determining the optimal graph. Hence, the development of a better algorithm is one of the important problems and we would like to discuss it in a future paper.

We showed the effectiveness of our method through the analysis of *Saccharomyces cerevisiae* gene expression data and evaluated the resulting network by comparing with biological knowledge. We construct the genetic network without using biological information. Nevertheless, the resulting network includes many important connections, which agree with the biological knowledge. Hence, we expect that our method can demonstrate its power in the analysis of a completely unknown system, like the human genome.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *in* B.N. Petrov, & F. Csaki, eds., Akademiai Kiado, Budapest, 267-281, 1973.

[2] H. Akaike. Likelihood and the Bayes procedure. *in* J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., Univ. Press, Valencia, 41-166, 1980.

[3] T. Akutsu, S. Miyano and S. Kuhara, S. Identification of Genetic Networks from a Small Number of Gene Expression Patterns Under the Boolean Network Model. *Proc. Pacific Symposium on Biocomputing*, **4**, 17-28, 1999.

[4] T. Akutsu, S. Miyano and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727-734, 2000.

[5] T. Akutsu, S. Miyano and S. Kuhara. Algorithms for Identifying Boolean Networks and Related Biological Networks Based on Matrix Multiplication and Fingerprint Function. *J. Comp. Biol.*, **7**, 331-344, 2000.

[6] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York, 1985.

[7] R. Cowell, A. Dawid, S. Lauritzen and D. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag New York, 1999.

[8] B. Daignan-Fornier and G.R. Fink. Coregulation of purine and histidine biosynthesis by the transcription activator BAS1 and BAS2. *Proc. Natl. Acad. Sci. USA*, **89**, 6746-6750, 1992.

[9] A.C. Davison. Approximate predictive likelihood. *Biometrika*, **73**, 323-332, 1986.

[10] C. De Boor. *A Practical Guide to Splines*. Springer-Verlag Berlin, 1978.

[11] V. Denis, H. Boucherie, C. Monribot and B. Daignan-Fornier. Role of the Myb-like protein Bas1p in *Saccharomyces cerevisiae*: a proteome analysis. *Mol. Microbiol.*, **30**, 556-566, 1998.

[12] N. Friedman and M. Goldszmidt. Discretizing Continuous Attributes While Learning Bayesian Networks. *Proc. 13'th International Conference on Machine Learning*, 157-165, 1996.

[13] N. Friedman and M. Goldszmidt. Learning Bayesian Networks with Local Structure. *Proc. Twelfth Conf. on Uncertainty in Artificial Intelligence*, 252-262, 1996.

[14] N. Friedman and M. Goldszmidt. Learning Bayesian Networks with Local Structure. *in* M.I. Jordan ed., Kluwer Academic Publisher, 1998.

[15] N. Friedman, M. Linial, I. Nachman and D. Pe'er. Using Bayesian Network to Analyze Expression Data. *J. Comp. Biol.*, **7**, 601-620, 2000.

[16] P.J. Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, 1994.

[17] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola and R.A. Young (2002). Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models. *Proc. Pacific Symposium on Biocomputing*, **7**, 437-449, 2002.

[18] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

[19] D. Heckerman. A tutorial on learning with Bayesian networks. *in* M.I. Jordan ed., Kluwer Academic Publisher, 1998.

[20] D. Heckerman and D. Geiger. Learning Bayesian networks: a unification for discrete and Gaussian domains. *Proc. Eleventh Conf. on Uncertainty in Artificial Intelligence*, 274-284, 1995.

[21] D. Heckerman, D. Geiger and D.M. Chickering. Learning Bayesian Networks: The combination of knowledge and statistical data. *Machine Learning*, **20**, 197-243, 1995.

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

[23] F.V. Jensen. *An introduction to Bayesian Networks*. University College London Press, 1996.

[24] S. Konishi. Statistical model evaluation and information criteria. *in* S. Ghosh ed., Marcel Dekker, 1999.

[25] S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, **83**, 875-890, 1996.

[26] H. Matsuno, A. Doi, M. Nagasaki and S. Miyano. Hybrid Petri Net representation of gene regulatory network. *Proc. Pacific Symposium on Biocomputing*, **5**, 338-349, 2000.

[27] H. Matsuno, A. Doi, Y. Hirata. and S. Miyano. XML Documentation of Biopathways and Their Simulations in Genomic Object Net. *Genome Informatics*, **12** 54-62, 2001.

[28] D. Pe'er, A. Regev, G. Elidan and N. Friedman. Inferring Subnetworks from Perturbed Expression Profiles. *Bioinformatics*, **17**, Suppl.1 (ISMB 2001), 215-224, 2001.

[29] R.J. Rolfes and A.G. Hinnebusch. Translation of the yeast transcriptional activator GCN4 is stimulated by purine limitation: implications for activation of the protein kinase GCN2. *Mol. Cell Biol.*, **13**, 5099-5111, 1993.

[30] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464, 1978.

[31] L. Tinerey and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82-86, 1986.