Combining Microarrays and Biological Knowledge for Estimating Gene Networks via Bayesian Networks

Seiya Imoto[1], Tomoyuki Higuchi[2], Takao Goto[1], Kousuke Tashiro[3], Satoru Kuhara[3] 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, takao, miyano}@ims.u-tokyo.ac.jp

[2]The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu,
Minato-ku, Tokyo, 106-8569, Japan
higuchi@ism.ac.jp

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

Abstract

We propose a statistical method for estimating a gene network based on Bayesian networks from microarray gene expression data together with biological knowledge including protein-protein interactions, protein-DNA interactions, binding site information, existing literature and so on. Unfortunately, microarray data do not contain enough information for constructing gene networks accurately in many cases. Our method adds biological knowledge to the estimation method of gene networks under a Bayesian statistical framework, and also controls the trade-off between microarray information and biological knowledge automatically. We conduct Monte Carlo simulations to show the effectiveness of the proposed method. We analyze Saccharomyces cerevisiae gene expression data as an application.

1. Introduction

In recent years, a large amount of gene expression data has been collected and estimating a gene network has become one of the central topics in the field of bioinformatics. Several methodologies have been proposed for constructing a gene network based on gene expression data, such as Boolean networks [1, 2, 32, 42], differential equation models [7, 10, 11, 32] and Bayesian networks [13, 14, 17, 18, 20, 22, 23, 37]. Main drawback for the gene network construction from microarray data is that while the gene network contains a large number of genes, the information contained in gene expression data is limited by the number of microarrays, their quality, the experimental design, noise, and measurement errors. Therefore, estimated gene networks contain some incorrect gene regulations, which cannot be evaluated from a biology viewpoint. In particular, the direction of gene regulation is difficult to decide using gene expression data only. Hence, the use of biological knowledge, including protein-protein and protein-DNA interactions [3, 5, 16, 21, 25], sequences of the binding site of the genes controlled by transcription regulators [31, 40, 47], literature and so on, are considered to be a key for microarray data analysis. The use of biological knowledge has previously received considerable attention for extracting more information from microarray data [4, 6, 18, 33, 36, 38, 41].

In this paper, we provide a general framework for combining microarray data and biological knowledge aimed at estimating a gene network by using a Bayesian network model. If the gene regulation mechanisms are completely known, we can model the gene network easily. However, many parts of the true gene network are still unknown and need to be estimated from data. Hence, it is necessary to construct a suitable criterion for evaluating estimated gene

networks in order to obtain an optimal network. While criteria proposed previously for evaluating a Bayesian network model only measure the closeness between a model and microarray data, we derive a criterion for selecting networks based on microarray data and biological knowledge. The proposed criterion is conducted by two components: One shows the fitness of the model to the microarray data and the other reflects biological knowledge, which is modeled under a probabilistic framework. Our proposed method automatically tunes the balance between the biological knowledge and microarray data based on our criterion and estimates a gene network from the combined data. In Section 2.1, we describe our statistical model for constructing gene networks and introduce a criterion for evaluating networks in Section 2.2. A statistical framework for representing biological knowledge is described in Section 2.3. In Section 2.4, we illustrate how to model various types of biological knowledge in practice. Monte Carlo simulations, in Section 3.1, are conducted to show the effectiveness of the proposed method. We apply our method to *Saccharomyces cerevisiae* gene expression data in Section 3.2.

2. Method for Estimating Gene Networks

2.1. Bayesian network and nonparametric heteroscedastic regression model

Bayesian networks [26] are a type of graphical models for capturing complex relationships among a large amount of random variables by the directed acyclic graph encoding the Markov assumption. In the context of Bayesian networks, a gene corresponds to a random variable shown as a node, while gene regulations are shown by directed edges. Thus gene interactions are modeled by the conditional distribution of each gene. We use Bayesian network and nonparametric heteroscedastic regression models [23] for constructing gene networks from microarray data.

Suppose that we have n sets of microarrays $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ of p genes, where $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ is a p dimensional gene expression vector obtained by ith microarray. Here, x_{ij} is an expression value of jth gene, denoted by gene$_j$, measured by ith microarray after required normalizations and transformation [39]. Ordinary, x_{ij} is given by $\log_2(R_{ij}/G_{ij})$, where R_{ij} and G_{ij} are normalized intensities of Cy5 and Cy3 for gene$_j$ measured by ith microarray. The interaction between gene$_j$ and its parents is modeled by the nonparametric additive regression model [19] with heterogeneous error variances

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

where $p_{ik}^{(j)}$ is the expression value of kth parent of gene$_j$ measured by ith microarray and ε_{ij} depends independently and normally on mean 0 and variance σ_{ij}^2. Here, $m_{jk}(\cdot)$ is a smooth function constructed by B-splines [9, 12, 24] of the form

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

where $\{b_{1k}^{(j)}(\cdot), ..., b_{M_{jk},k}^{(j)}(\cdot)\}$ is a prescribed set of B-splines and $\gamma_{mk}^{(j)}$ are parameters. Hence, a Bayesian network and nonparametric heteroscedastic regression model can be represented as

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

for $i = 1, ..., n$, where $\boldsymbol{\theta}_G$ is a parameter vector and $f_j(x_{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\theta}_j)$ is a density of Gaussian distribution with mean $m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)})$ and variance σ_{ij}^2. If gene$_j$ has no parent genes, we use μ_j and σ_j^2 instead of $m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)})$ and σ_{ij}^2, respectively.

This model has several advantages. Unlike Boolean networks and discrete Bayesian networks [13, 14, 17, 18, 20, 37], no discretization of gene expression data, which leads to information loss, is required. Second, even nonlinear relationships between genes are automatically extracted based on gene expression data.

2.2. Criterion for evaluating networks

Some gene networks are partially known, but many mechanisms of gene regulations are still unknown. Therefore we need to estimate unknown structures of the gene network from the data. Hence, the construction of a suitable criterion for measuring the closeness between an estimated gene network and the true one is an essential problem for statistical gene network modeling. Following the result of Imoto *et al.* [23], a criterion for evaluating an estimated gene network can be derived from Bayes approach. At first, we briefly introduce the derivation of their criterion. We then explain how extend their criterion for combining microarray data and biological knowledge.

When we construct a gene network G by using a Bayesian network model, the posterior probability of the network is obtained as the product of prior probability of the network $\pi(G)$ and the marginal likelihood divided by the normalizing constant. After dropping the normalizing constant, the posterior probability of the network is proportional to

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

where $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ is a prior distribution on the parameter vector $\boldsymbol{\theta}_G$ with hyperparameter vector $\boldsymbol{\lambda}$ satisfying $\log \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(n)$. The essential problem for constructing a criterion based on the posterior probability of the network is how to compute the marginal likelihood given by a high dimensional integral. Imoto *et al.* [23] used the Laplace approximation for integrals [8, 30, 45] and derived a criterion, named BNRC_{hetero} (Bayesian network and Nonparametric heteroscedastic Regression Criterion), of the form

$$\begin{aligned}\mathrm{BNRC}_{hetero}(G) &= -2\log \pi(G) \\ &\quad + \log\left|\frac{n}{2\pi} J_\lambda(\hat{\boldsymbol{\theta}}_G)\right| - 2n l_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}),\end{aligned}$$

where

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

and $\hat{\boldsymbol{\theta}}_G$ is the mode of $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})$.

Suppose that the prior distribution $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ is factorized as

$$\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \prod_{j,k} \pi_{jk}(\boldsymbol{\gamma}_{jk}|\boldsymbol{\lambda}_{jk}),$$

where $\boldsymbol{\gamma}_{jk} = (\gamma_{1k}^{(j)}, ..., \gamma_{M_{jk},k}^{(j)})^T$ is a parameter vector and λ_{jk} is a hyperparameter. We use a singular M_{jk} variate normal distribution as the prior distribution on $\boldsymbol{\gamma}_{jk}$,

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

where K_{jk} is an $M_{jk} \times M_{jk}$ symmetric positive semidefinite matrix satisfying $\boldsymbol{\gamma}_{jk}^T K_{jk}\boldsymbol{\gamma}_{jk} = \sum_{\alpha=3}^{M_{jk}} (\gamma_{\alpha k}^{(j)} - 2\gamma_{\alpha-1,k}^{(j)} + \gamma_{\alpha-2,k}^{(j)})^2$. Then we have the decomposition $\mathrm{BNRC}_{hetero} = -2\log \pi(G) + \sum_{j=1}^{p} \mathrm{BNRC}_{hetero}^{(j)}$. Here $\mathrm{BNRC}_{hetero}^{(j)}$ is a score for gene$_j$ and given by

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

where $w_{ij},\ i = 1, ..., n$ are weights of the heterogeneous error variance $\sigma_{ij}^2 = w_{ij}^{-1}\sigma_j^2$ and $\Lambda_j = B_{jk}^T W_j B_{jk} + n\beta_{jk}K_{jk}$ with $B_{jk} = (\boldsymbol{b}_{jk}(p_{1k}^{(j)}), ..., \boldsymbol{b}_{jk}(p_{nk}^{(j)}))^T$, $\boldsymbol{b}_{jk}(p_{ik}^{(j)}) = (b_{1k}^{(j)}(p_{ik}^{(j)}), ..., b_{M_{jk},k}^{(j)}(p_{ik}^{(j)}))^T$, $W_j = \mathrm{diag}(w_{1j}, ..., w_{nj})$ and $\beta_{jk} = \sigma_j^2\lambda_{jk}$. The details of the parameter estimation are described in Imoto *et al.* [23].

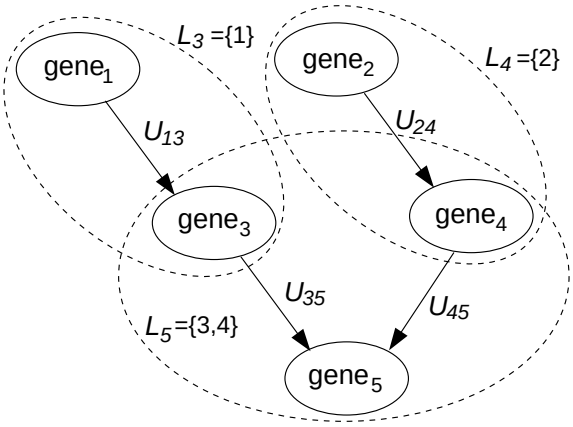


Figure 1. A gene network and its energy. The index sets L_3, L_4 and L_5 are illustrated and L_1 and L_2 are defined by empty sets. The local energies are $E_3 = U_{13}$, $E_4 = U_{24}$ and $E_5 = U_{35} + U_{45}$. The total energy of this network is $E = E_3 + E_4 + E_5 = U_{13} + U_{24} + U_{35} + U_{45}$.

2.3. Adding biological knowledge

The criterion $\mathrm{BNRC}_{hetero}(G)$, introduced in the previous section, contains two quantities: the prior probability $\pi(G)$ of the network, and the marginal likelihood of the data. The marginal likelihood shows the fitness of the model to the microarray data. The biological knowledge can then be added into the prior probability of the network $\pi(G)$.

Let U_{ij} be the interaction energy of the edge from gene$_i$ to gene$_j$ and let U_{ij} be categorized into I values, $H_1, ..., H_I$, based on biological knowledge. For example, if we know *a priori* gene$_i$ regulates gene$_j$, we set $U_{ij} = H_1$. However, if we do not know whether gene$_k$ regulates gene$_l$ or not, we set $U_{kl} = H_2$. Note that $0 < H_1 < H_2$. The

total energy of the network G can then be defined as

$$E(G) = \sum_{\{i,j\} \in G} U_{ij},$$

where the sum is taken over the existing edges in the network G. Under the Bayesian network framework, the total energy can be decomposed into the sum of the local energies

$$E(G) = \sum_{j=1}^{p} \sum_{i \in L_j} U_{ij} = \sum_{j=1}^{p} E_j, \quad (1)$$

where L_j is an index set of parents of gene$_j$ and $E_j = \sum_{i \in L_j} U_{ij}$ is a local energy defined by gene$_j$ and its parents. Figure 1 shows an example of a gene network and its energy.

The probability of a network G, $\pi(G)$, is naturally modeled by the Gibbs distribution [15]

$$\pi(G) = Z^{-1} \exp\{-\zeta E(G)\}, \quad (2)$$

where ζ (> 0) is a hyperparameter and Z is a normalizing constant called the partition function

$$Z = \sum_{G \in \mathcal{G}} \exp\{-\zeta E(G)\}.$$

Here $\mathcal{G}$ is the set of possible networks. By replacing $\zeta H_1, ..., \zeta H_I$ with $\zeta_1, ..., \zeta_I$, respectively, the normalizing constant Z is a function of $\zeta_1, ..., \zeta_I$. We call ζ_j an inverse normalized temperature. By substituting (1) into (2), we have

$$\begin{aligned} \pi(G) &= Z^{-1} \prod_{j=1}^{p} \exp\{-\zeta E_j\} \\ &= Z^{-1} \prod_{j=1}^{p} \prod_{i \in L_j} \exp(-\zeta_{\alpha(i,j)}), \end{aligned}$$

with $\alpha(i,j) = k$ for $U_{ij} = H_k$. Hence, by adding biological knowledge into the prior probability of the network, BNRC$_{hetero}$ can be rewritten as

$$\begin{aligned} \mathrm{BNRC}_{hetero}(G, \zeta_1, ..., \zeta_I) &= 2 \log Z \\ &+ \sum_{j=1}^{p} \{2 \sum_{i \in L_j} \zeta_{\alpha(i,j)} + \mathrm{BNRC}^{(j)}_{hetero}\}. \end{aligned} \quad (3)$$

We can choose an optimal network under the given $\zeta_1, ..., \zeta_I$. Also the optimal values of $\zeta_1, ..., \zeta_I$ are obtained as the minimizer of (3). Therefore, we can represent an algorithm for estimating a gene network from microarray data and biological knowledge as follows:

Step1: Set the values $\zeta_1, ..., \zeta_I$.
Step2: Estimate a gene network by minimizing $\mathrm{BNRC}_{hetero}(G)$ under the given $\zeta_1, ..., \zeta_I$.
Step3: Repeat Step1 and Step2 against the candidate values of $\zeta_1, ..., \zeta_I$.
Step4: An optimal gene network is obtained from the candidate networks obtained in Step3.

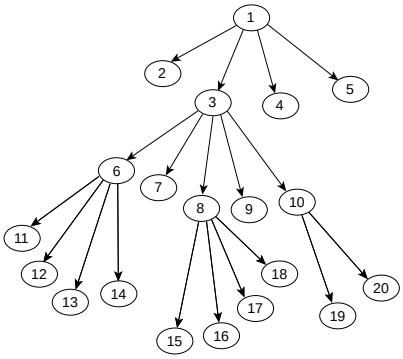
In Step2, we use the greedy hill-climbing algorithm for learning networks. The details are shown in Imoto *et al.* [23]. Note that the proposed prior probability of the network can be used for other types of Bayesian network models, such as discrete Bayesian networks and dynamic Bayesian networks [29, 34, 36, 43].

The computation of partition function, Z, is intractable even for moderately sized gene networks. To avoid this problem, we compute upper and lower bounds of the partial function and use them for choosing the optimal values of $\zeta_1, ..., \zeta_I$. An upper bound is obtained by directed graphs, which are allowed to contain cyclic graphs. Thus the true value of the partition function is not greater than the upper bound. A lower bound is computed by multi-level directed graphs with following assumptions: (A1) There is one top gene and (A2) Genes at the same level have a common parent gene that is located on one upper level of them. We also consider joined graphs of some multi-level directed graphs satisfying (A1) and (A2). Since the number of possible graphs is much larger than those included in the computation, the true value of the partition function should be greater than the lower bound. Since the optimization of the network structure for fixed $\zeta_1, ..., \zeta_I$ does not depend on the value of the partition function, our method works well in practice. Of course, when the number of genes is small, we can perform an exhaustive search and compute the partition function completely. However, we think that the development of an effective algorithm to enumerate all possible networks or approximate the partition function is an important problem.

2.4. Prior design for various biological knowledge

In this subsection, we show some examples of biological knowledge and how to include them into the prior probability in practice. We consider using two values ζ_1 and ζ_2 satisfying $0 < \zeta_1 < \zeta_2$ for representing biological knowledge. Basically, we allocate ζ_1 to a known relationship and ζ_2 otherwise. The prior information can be summarized as a $p \times p$ matrix U whose (i,j) element, u_{ij}, corresponds to ζ_1 or ζ_2.

Protein-protein interactions
The number of known protein-protein interactions is rapidly increasing and kept in some public databases such



(a)

$$g_1 = \varepsilon_1,\ g_2 = .7g_1 + \varepsilon_2\ g_5 = .7g_1 + \varepsilon_5,\ g_{10} = 1/\{1+\exp(-4g_3)\} + \varepsilon_{10}$$

$$g_3 = \begin{cases} -1 + \varepsilon_3\ (g_1 \le -.5) \\ g_1 + \varepsilon_3\ (|g_1| < .5) \\ 1 + \varepsilon_3\ (g_1 \ge .5) \end{cases} \quad g_6 = \begin{cases} .8g_3 + \varepsilon_6\ (g_3 \le -1) \\ (g_3+1)^{1.5} + \varepsilon_6\ (-1 < |g_3| < 0) \\ 1 + \varepsilon_6\ (g_3 \ge 1) \end{cases}$$

$$g_4 = \begin{cases} .4g_1 + 1 + \varepsilon_4\ (|g_1| \le .3) \\ (g_1+1)^2 + \varepsilon_4\ (|g_1| < .3) \end{cases} \quad g_8 = \begin{cases} .2g_3 - 1 + \varepsilon_8\ (g_3 \le .2) \\ 1.4g_3 + \varepsilon_8\ (g_3 > .2) \end{cases}$$

$$g_{11} = .7g_6 + \varepsilon_{11},\ g_{14} = .7g_6 + \varepsilon_{14},\ g_{15} = 1/\{1+\exp(-4g_8)\} + \varepsilon_{15}$$

$$g_9 = \begin{cases} .4g_3 + 1 + \varepsilon_9\ (|g_3| \le .3) \\ (g_3+1)^{1.2} + \varepsilon_9\ (|g_3| < .3) \end{cases} \quad g_{13} = \begin{cases} .4g_6 + 1 + \varepsilon_{13}\ (|g_6| \le .3) \\ (g_6+1)^2 + \varepsilon_{13}\ (|g_6| < .3) \end{cases}$$

$$g_{12} = \begin{cases} -1 + \varepsilon_{12}\ (g_6 < -.5) \\ g_6 + \varepsilon_{12}\ (|g_6| \le .5) \\ 1 + \varepsilon_{12}\ (|g_6| > .5) \end{cases} \quad \begin{aligned} g_{16} &= .8g_8 + \varepsilon_{16} \\ g_{19} &= 1/\{1+\exp(-4g_{10})\} + \varepsilon_{19} \\ g_{20} &= 1.1g_{10} + \varepsilon_{20} \end{aligned}$$

$$g_{17} = \begin{cases} .2g_8 - 1 + \varepsilon_{17}\ (g_8 \le .2) \\ 1.4g_8 + \varepsilon_{17}\ (g_8 > .2) \end{cases} \quad g_{18} = \begin{cases} .4g_8 + 1\ (|g_8| > .3 \\ (g_8+1)^{1.2}\ (g_8 \le .3) \end{cases}$$

(b)

Figure 2. Artificial gene network and functional structures between nodes.

as GRID [16] and BIND [3, 5]. Protein-protein interactions show at least two proteins form a complex. Therefore, representing protein-protein interactions by a directed graph is not suitable. However, they can be included in our method. If we know gene$_i$ and gene$_j$ create a protein-protein interaction, we set $u_{ij} = u_{ji} = \zeta_1$. In such a case, we will decide whether we make a virtual node corresponding to a protein complex theoretically [35].

Protein-DNA interactions
Protein-DNA interactions show gene regulations by transcription factors and can be modeled more easily than protein-protein interactions. When gene$_i$ is a transcription regulator and controls gene$_j$, we set $u_{ij} = \zeta_1$ and $u_{ji} = \zeta_2$.

Sequences
Genes that are controlled by a transcription regulator might have a consensus motif in their promoter DNA sequences. If gene$_{j_1}$,...,gene$_{j_n}$ have a consensus motif and are controlled by gene$_i$, we set $u_{ij_1} = \cdots = u_{ij_n} = \zeta_1$ and $u_{j_1 i} = \cdots = u_{j_n i} = \zeta_2$. Previously, consensus motifs were often used for the evaluation of estimated gene networks from a biological viewpoint. This information, however, can be introduced directly into our method. One straightforward way is the use of known regulatory motifs kept in public databases such as SCPD [40] and YTF [47]. As for an advanced method, Tamada *et al.* [44] proposed a method for simultaneously estimating a gene network and detecting regulatory motifs based on our method, and succeeded in estimating an accurate gene network and detecting a true regulatory motif.

Gene networks and pathways
The information of gene networks can be introduced directly into our method by transforming the prescribed network structures into the matrix U. We can then estimate a gene network based on U and microarray data. Our method also can use gene networks estimated by other techniques such as boolean networks, differential equation models, and so on. Also, some databases, such as KEGG [28], contain several known gene networks and pathways. This information can be used similarly.

Literature
Some research has been performed to extract information from a huge amount of literature [27]. Literature contain various kinds of information including biological knowledge described above. So we can model literature information in the same way.

3. Computational Experiments

3.1. Monte Carlo simulations

Before analyzing real gene expression data, we perform Monte Carlo simulations to examine the properties of the proposed method. We assume an artificial network with 20 nodes shown in Figure 2 (a). The functional relationships between nodes are listed in Figure 2 (b). A network will be rebuilt from simulated data consisting of 50 or 100 observations, which corresponds to 50 or 100 microarrays. As for the biological knowledge, we tried the following situations: **(Case 1)** we know some gene regulations (100%, 75%, 50% or 25% out of 19 edges shown in Figure 2 (a)) and **(Case 2)** we know some gene regulations, but some (1, 2, or 3) incorrect edges are kept in the database. The candidate values of ζ_1 and ζ_2 are $\{0.5,\ 1.0\}$ and $\{\zeta_1,\ 2.5,\ 5.0,\ 7.5,\ 10.0\}$, respectively.

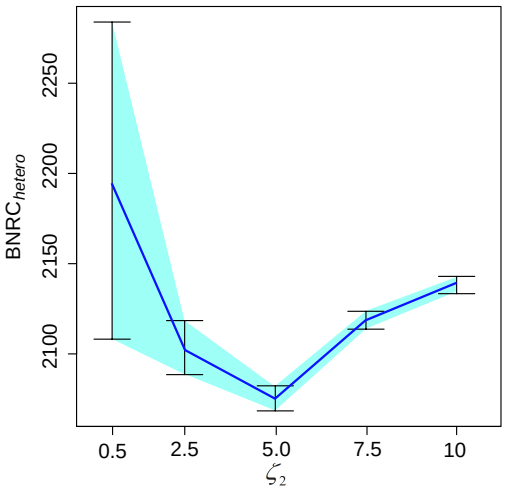


Figure 3. The behavior of BNRC$_{hetero}$ when ζ_1 = 0.5. We can find out the optimal inverse normalized temperature ζ_2 is 5.0.

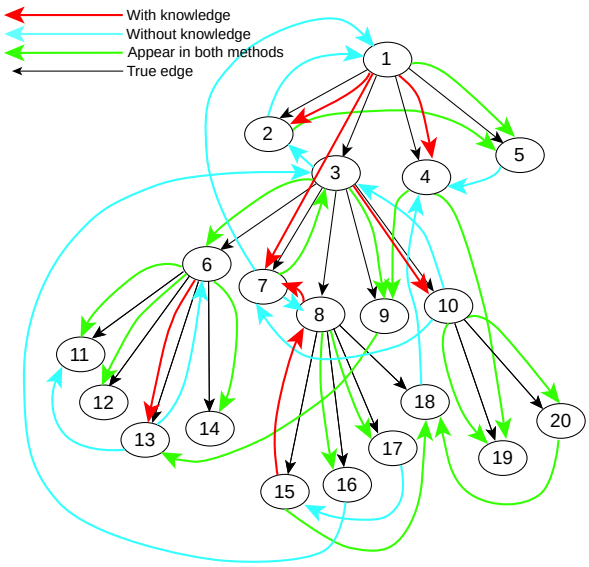


Figure 4. An example of resulting networks based on 100 samples. We used ζ_1 = 0.5 and ζ_2 = 5.0 that are selected by our criterion (see Figure 3).

Figure 4 shows two estimated networks: One is estimated by 100 observations (microarrays) alone. We use $\zeta_1 = \zeta_2 = 0.5$, i.e. we did not use any knowledge (we denote this network by N_0 for convenience). The other is estimated by 100 observations and prior information of 75% gene regulations, i.e. we know 14 correct relations out of the all 19 correct edges (we denote this network by N_1). Edges appearing in both networks are colored green, while edges appearing in N_0 or N_1 only are colored blue and red, respectively. By adding prior knowledge, it is clear that we succeeded in reducing the number of false positives. We also find additional four correct relationships. Figure 3 shows the behavior of BNRC$_{hetero}$ when $\zeta_1 = 0.5$. We find that the optimal value of ζ_2 is 5.0. From the Monte Carlo simulations, we observed that ζ_2 can be selected by using middle values (depicted by a blue line) of upper and lower bounds or upper bounds in practice. For the selection of ζ_1, we use the middle value of the upper and lower bounds of the score of our criterion.

The results of the Monte Carlo simulations are summarized as follows:
In **(Case 1)**, we obtained networks more accurately as long as we add correct knowledge. We observed that the number of false positives decreased drastically. We presume the reason is the nature of directed acyclic graphs. Since a Bayesian network model is a directed acyclic graph, one incorrect estimate may affect the relations in its neighborhood. However, by adding some correct knowledge, we can restrict the search space of the Bayesian network model learning effectively.
In **(Case 2)**, the results depend on the type of incorrect knowledge.
(i) If we use misdirected relations, e.g. gene$_8$ $\rightarrow$ gene$_3$, as prior knowledge, serious problems occur. Since microarray data to some degree support the misdirected relations, they tend to receive a better criterion score.
(ii) If we add indirect relations such as gene$_1$ $\rightarrow$ gene$_8$, we observed that our method controlled the balance between this prior information and microarray data and could decide whether the prior relation is true.
(iii) If irrelevant relations such as gene$_{20}$ $\rightarrow$ gene$_5$ are added as prior information, we observed that our method could reject these prior information, because, the microarray data do not support these relations.

3.2. Example using experimental data

In this subsection, we demonstrate our method by analyzing *Saccharomyces cerevisiae* gene expression data obtained by disrupting 100 genes, which are almost all transcription factors. We focus on five genes, *MCM1*, *SWI5*, *ACE2*, *SNF2* and *STE12* (see Table 1) and extract genes that are regulated by these 5 genes from the Yeast Proteome Database [46]. Thus, we construct a prior network shown in Figure 5, based on the database information. We include the prior network in our Bayesian network estimation method.

MCM1 :	transcription factor of the MADS box family
	MET14, *CDC6*, *MET2*, *CDC5*, *MET6*, *SIC1*, *STE6*, *CLN2*, *PCL2*, *STE2*, *ACE2*, *MET16*, *MET3*, *MET4*, *CAR1*, *SWI5*, *PCL9*, *CLB1*, *MET17*, *EGT2*, *ARG5,6*, *PMA1*, *RME1*, *CLB2*
SWI5 :	transcription factor
	CDC6, *SIC1*, *CLN2*, *PCL2*, *PCL9*, *EGT2*, *RME1*, *CTS1*, *HO*
ACE2 :	metallothionein expression activator
	CLN2, *EGT2*, *HO*, *CTS1*, *RME1*
SNF2 :	component of SWI/SNF global transcription activator complex
	CTS1, *HO*
STE12 :	transcriptional activator
	STE6, *FAR1*, *KAR3*, *SST2*, *FUS1*, *STE2*, *BAR1*, *AGA1*, *AFR1*, *CIK1*

Table 1. Five transcription factors and their regulating genes.

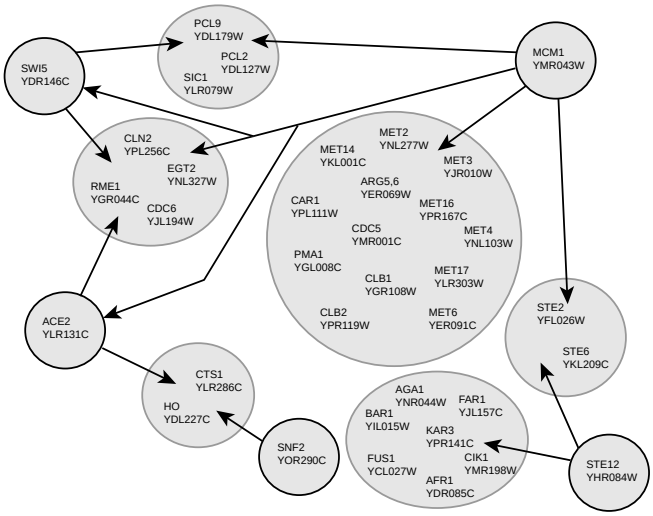


Figure 5. Prior knowledge network. The genes that are in each shadowed circle are regulated by the parent genes.

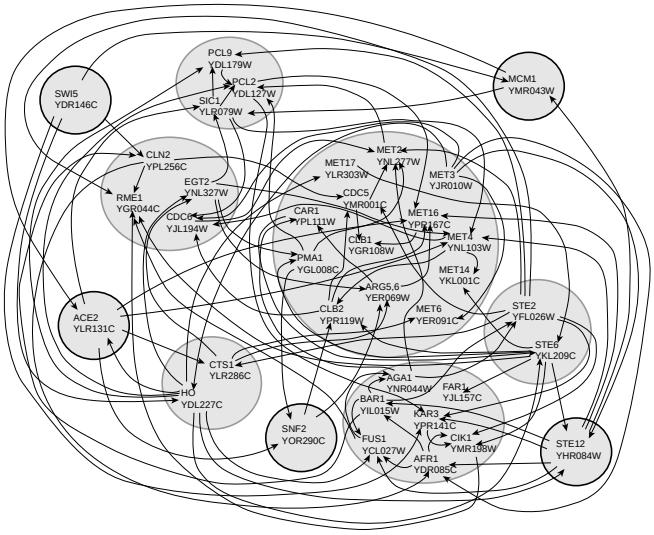


Figure 6. Resulting network based on microarray only.

That is, the purpose of this analysis is to estimate the gene network containing above 36 genes from microarray data together with the prior network. Figure 6 shows the estimated gene network using microarray data only. There are many non-prior edges and many of them are probably false positives. In addition, we find three misdirected relations: "*SWI5* $\rightarrow$ *MCM1*", "*HO* $\rightarrow$ *ACE2*" and "*STE6* $\rightarrow$ *STE12*". By adding the prior network, we obtain the gene network shown in Figure 8. As for the inverse normalized temperatures ζ_1 and ζ_2, we set $\zeta_1 = 0.5$ and choose the optimal value of ζ_2. We also estimated a gene network based on $\zeta_1 = 1$ and found the results described below to be essentially unchanged.

Figure 7 shows the behavior of BNRC_{hetero} with respect to ζ_2. We find that the optimal value of ζ_2 is 2.5. Figure 8 shows the resulting network based on microarray data and the biological knowledge represented by the prior network in Figure 5. We show the edges that correspond to the prior knowledge in black. The edges between genes that are regulated by the same transcription factor in the prior network are shown in blue. The red edges do not correspond to the prior knowledge. In particular, we find that the relationships around *MCM1* improve drastically. The network based on microarray only (Figure 6) indicates that only *SIC1* and *ACE2* are regulated by *MCM1*. Note that the underlined genes correspond to the prior network information. After adding the prior knowledge and optimizing the inverse normalized temperatures, we find that 10 genes out of 24 genes that are listed as co-regulated genes of *MCM1* in Table 1 are extracted. Also, the relationships around *STE12*

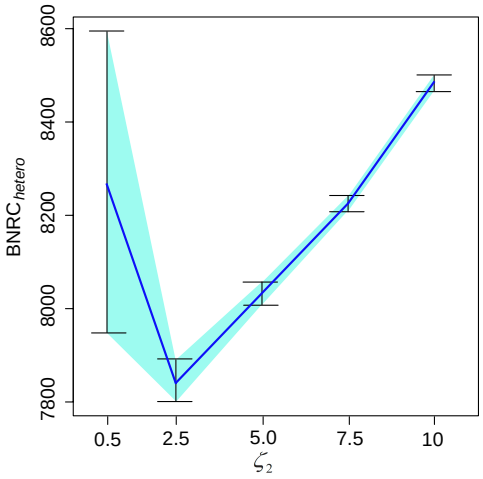


Figure 7. Optimization of ζ_2. We can find out that the optimal value of ζ_2 is 2.5.

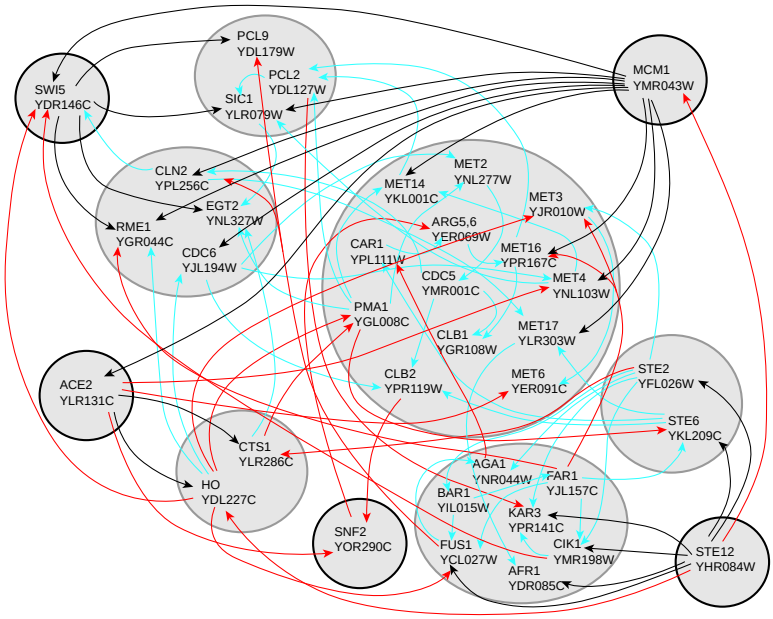


Figure 8. Resulting network based on microarray data and biological knowledge. The inverse normalized temperatures are selected by our criterion (ζ_1=0.5, ζ_2=2.5).

become clearer. Before adding prior knowledge, the estimated network in Figure 6 suggests *FUS1*, *AFR1*, *KAR3*, *BAR1*, *MET4*, *MET16* and *MCM1* are regulated by *STE12*, while *STE12* is controlled by *HO*, *STE6* and *MET3*. On the other hand, the network in Figure 8 shows that *STE12* regulates *FUS1*, *AFR1*, *KAR3*, *CIK1*, *STE2*, *STE6*, *HO* and *MCM1*. Note that the three misdirected relations described above are corrected in Figure 8. The difference between the inverse normalized temperatures $\zeta_1 = 0.5$ and $\zeta_2 = 2.5$ is small, because the score of the criterion is added as $2\zeta_1$ or $2\zeta_2$, when we add an edge that is listed or not listed in the prior network, respectively. Therefore, microarray data contain this information and we succeeded in extracting this information with the slight help of the prior network.

We optimized the inverse normalized temperature ζ_2 based on the proposed criterion. From the network based on the optimal inverse normalized temperatures, we can find the gap between microarray data and biological knowledge. By comparing Figure 6 with Figure 8, we find that the microarray data reflect the relationship between seven genes (*CLN2*, *RME1*, *CDC6*, *EGT2*, *PCL2*, *PCL9* and *SIC1*) and two transcription factors (*MCM1* and *SWI5*). On the other hand, we find that there are somewhat large differences between microarray data and the prior network for the relationship between *MCM1* and the thirteen genes that are in the biggest circle.

4. Discussion

In this paper we proposed a general framework for combining microarray data and biological knowledge aimed at estimating a gene network. An advantage of our method is the balance between microarray information and biological knowledge is optimized by the proposed criterion. By adding biological knowledge into our Bayesian network estimation method, we succeeded in extracting more information from microarray data and estimating the gene network more accurately. We believe that the combination of microarray data and biological knowledge gives a new perspective for understanding the systems of living creatures.

We consider the following problems as our future works: (1) In the real application, we demonstrated how to use the gene network that is obtained biologically as a prior knowledge. There are various types of biological knowledge we listed in Section 2.4. It is a very important problem how to use such knowledge together with microarray data in practice. (2) From biological knowledge, we deterministically decided the category to which edges belong, e.g. $u_{11} = \zeta_1$, $u_{12} = \zeta_2$, and so on. However, biological knowledge contains some errors. In fact, u_{ij} can be viewed as a random variable, and a statistical model can be constructed for u_{ij}. In that sense, our method can be extended as a Bayesian network estimation method with a self-repairing database mechanism. We would like to investigate these problems in a future paper.

References

[1] T. Akutsu, S. Miyano and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pacific Symposium on Biocomputing*, **4**, 17-28, 1999.

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

[3] G.D. Bader, I. Donaldson, C. Wolting, B.F.F. Ouellette, T. Pawson and C.W.V. Hogue. BIND-The biomolecular interaction network database. *Nucleic Acids Research*, **29**, 242-245, 2001.

[4] H. Bannai, S. Inenaga, A. Shinohara, M. Takeda and S. Miyano. A string pattern regression algorithm and its application to pattern discovery in long introns. *Genome Informatics*, **13**, 3-11, 2002.

[5] BIND http://www.blueprint.org/

[6] H.J. Bussemaker, H. Li and E.D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, **27**, 167-171, 2001.

[7] T. Chen, H. He and G. Church. Modeling gene expression with differential equations. *Pacific Symposium on Biocomputing*, **4**, 29-40, 1999.

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

[9] C. de Boor. *A Practical Guide to Splines*. Springer, Berlin. 1978.

[10] M.J.L. de Hoon, S. Imoto and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data using differential equations. *Proc. 5th International Conference on Discovery Science*, Lecture Note in Artificial Intelligence, **2534**, Springer-Verlag, 267-274, 2002.

[11] M.J.L. de Hoon, S. Imoto, K. Kobayashi, N. Ogasawara and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pacific Symposium on Biocomputing*, **8**, 17-28, 2003.

[12] P.H.C. Eilers and B. Marx. Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, **11**, 89-121, 1996.

[13] N. Friedman and M. Goldszmidt. *Learning Bayesian networks with local structure*. in M.I. Jordan ed., Kluwer Academic Publisher, 421-459, 1998.

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

[15] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restorations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741, 1984.

[16] GRID http://biodata.mshri.on.ca/grid/servlet/Index

[17] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola and R.A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing*, **6**, 422-433, 2001.

[18] A.J. Hartemink, D.K. Gifford, T.S. Jaakkola and R.A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. *Pacific Symposium on Biocomputing*, **7**, 437-449, 2002.

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

[20] D. Heckerman. *A tutorial on learning with Bayesian networks*. in M.I. Jordan ed., Kluwer Academic Publisher, 301-354, 1998.

[21] T. Ideker, O. Ozier, B. Schwikowski and A.F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (ISMB 2002), S233-S240, 2002.

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

[23] S. Imoto, S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara and S. Miyano. Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *Journal of Bioinformatics and Computational Biology*, in press. (Preliminary version has appeared in *Proc. 1st IEEE Computer Society Bioinformatics Conference*, 219-227, 2002).

[24] S. Imoto and S. Konishi. Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Ann. Inst. Statist. Math.*, in press.

[25] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA*, **97**, 4569-4574, 2001.

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

[27] T.-K. Jenssen, A. L greid. J. Komorowski and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, **28**, 21-28.

[28] KEGG http://www.genome.ad.jp/kegg/kegg2.html

[29] S. Kim, S. Imoto and S. Miyano. Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. *Proc. 1st International Workshop on Computational Methods in Systems Biology*, Lecture Note in Computer Science, **2602**, Springer-Verlag, 104-113, 2003.

[30] S. Konishi, T. Ando and S. Imoto. Bayesian information criteria and smoothing parameter selection in radial basis function networks. submitted for publication.

[31] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B.Ren, J.J. Wyrick, J-B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford and R.A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799-804, 2002

[32] Y. Maki, D. Tominaga, M. Okamoto, S. Watanabe and Y. Eguchi. Development of a system for the inference of large scale genetic networks. *Pacific Symposium on Biocomputing*, **6**, 446-458, 2001.

[33] D.R. Masys. Linking microarray data to the literature. *Nature Genetics*, **28**, 9-10, 2001.

[34] K. Murphy and S. Mian. Modelling gene expression data using dynamic Bayesian networks. *Technical report, Computer Science Division, University of California, Berkeley, CA*. 1999.

[35] N. Nariai, S. Kim, S. Imoto and S. Miyano. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. under preparation.

[36] I.M. Ong, J.D. Glasner and D. Page. Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics*, **18** (ISMB2002), S241-S248, 2002.

[37] D. Pe'er, A. Regev, G. Elidan and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17** (ISMB 2001), S215-S224, 2001.

[38] Y. Pilpel, P. Sudarsanam and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, **29**, 153-9, 2001.

[39] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, **32**, 496-501, 2002.

[40] SCPD http://cgsigma.cshl.org/jian/

[41] E. Segal, Y. Barash, I. Simon, N. Friedman and D. Koller. From promoter sequence to expression: a probabilistic framework. *Proc. 6th Annual International Conference on Research in Computational Molecular Biology* (RECOMB 2002), 263-272, 2002.

[42] I. Shmulevich, E.R. Dougherty, S. Kim and W. Zhang. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**, 261-274, 2002.

[43] V.A. Smith, E.D. Jarvis and A.J. Hartemink. Evaluating functional network inference using simulations of complex biological systems. *Bioinformatics*, **18** (ISMB 2002), S216-S224, 2002.

[44] Y. Tamada, S. Kim, H. Bannai, S. Imoto K. Tashiro, S. Kuhara and S. Miyano. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, (ECCB 2003), in press.

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

[46] YPD http://www.incyte.com/sequence/proteome/databases/YPD.shtml

[47] YTF http://biochemie.web.med.uni-muenchen.de/YTFD/