BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii227–ii236
DOI: 10.1093/bioinformatics/btg1082

Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection

Yoshinori Tamada[1,*,†], SunYong Kim[1,†], Hideo Bannai[1], Seiya Imoto[1], Kousuke Tashiro[2], Satoru Kuhara[2] and Satoru Miyano[1]

[1]Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan and [2]Graduate School of Genetic Resource Technology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan

Received on March 17, 2003; accepted on June 9, 2003

ABSTRACT

We present a statistical method for estimating gene networks and detecting promoter elements simultaneously. When estimating a network from gene expression data alone, a common problem is that the number of microarrays is limited compared to the number of variables in the network model, making accurate estimation a difficult task. Our method overcomes this problem by integrating the microarray gene expression data and the DNA sequence information into a Bayesian network model. The basic idea of our method is that, if a parent gene is a transcription factor, its children may share a consensus motif in their promoter regions of the DNA sequences. Our method detects consensus motifs based on the structure of the estimated network, then re-estimates the network using the result of the motif detection. We continue this iteration until the network becomes stable. To show the effectiveness of our method, we conducted Monte Carlo simulations and applied our method to *Saccharomyces cerevisiae* data as a real application.

Contact: tamada@ims.u-tokyo.ac.jp

INTRODUCTION

Constructing gene networks from microarray gene expression data is becoming an important challenge in the post-genomic era. A *gene network*, or gene regulatory network, is a model which represents regulations between genes using a directed graph. In gene networks, nodes indicate genes and edges represent regulations between genes (e.g. activation or suppression). Several methods have been proposed for estimating gene networks from microarray data using mathematical models such as Boolean networks (Akutsu *et al.*, 1999, 2000a,b, 2003; Maki *et al.*, 2001; Shmulevich *et al.*, 2002), differential equations (Chen *et al.*, 1999; De Hoon *et al.*, 2003), and Bayesian networks (Friedman and Goldszmidt, 1998; Friedman *et al.*, 2000; Imoto *et al.*, 2002). Although these methods succeed in constructing networks where genes known to be biologically related come close together, it is difficult to determine the correct direction of the edges, as well as whether or not the relation of genes is direct or indirect. This is especially true when using disruptant microarray data (as opposed to time series microarray experiments which contain information concerning time dependencies) (e.g. see Fig. 5 in Imoto *et al.* (2003a)).

The drawbacks of the previous methods are mainly caused by the limited number of microarrays. From a statistical point of view, the number of samples (microarrays) is always insufficient to estimate accurate networks as opposed to the number of variables (genes) in the model. Theoretically, this problem can be solved by using more microarrays, but this solution is unrealistic because of the cost incurred in producing a sufficient number of microarray data.

To overcome these problems, we have developed a statistical method for estimating gene networks by utilizing DNA sequences and microarray data. The basic idea is as follows: The regulation of genes is known to be realized by transcription factors (TFs), which are important subsets of proteins that transcribe mRNAs from DNAs. Genes that a specific TF regulates, contain a binding consensus motif called the transcription factor binding site, located in the upstream regions of the genes. Suppose that a gene g in the network is a transcription factor. If the children of g are directly regulated by g, then they may share a consen-

[*]To whom correspondence should be addressed. Current affiliation: Bioinformatics Center, Institute for Chemical Research, Kyoto University.

[†] These authors contributed equally to this work.

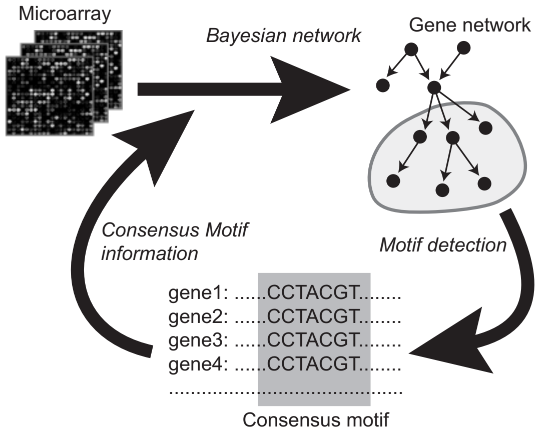

Microarray
Bayesian network
Gene network
Consensus Motif information
Motif detection
gene1: ......CCTACGT........
gene2: ......CCTACGT........
gene3: ......CCTACGT........
gene4: ......CCTACGT........
Consensus motif

Fig. 1. Conceptual view of our proposed method.

sus motif in their upstream DNA sequences. By detecting a consensus motif from a set of genes which have been selected based on the structure of the network, we can correct the network by repairing mis-directed edges and/or adding direct edges from g, based on the existence of the motif.

Figure 1 represents the conceptual explanation of our method. First, we estimate a gene network from microarray data alone using a Bayesian network model (Imoto *et al.*, 2002, 2003a). Based on the network structure, we then focus on several genes which are regarded as transcription factor candidates in the estimated network, and define sets of genes that may be co-regulated by each candidate. A motif detection method (Bannai *et al.*, 2002, 2003) is then performed for detecting a consensus motif from each set of possibly co-regulated genes. After the motif detection, we revise the structure of the network based on the motif information and embed this information into a Bayesian network estimation method as a prior probability of the network. The network is estimated again using both microarray data and the motif information this time. This iterative procedure, the motif detection and the network re-estimation, is repeated until the structure of the network does not change considerably.

To evaluate our method, we first conducted Monte Carlo simulations. We designed an artificial network and generated pseudo microarray data and pseudo DNA sequences. We compared the estimated network by our method with one estimated by only microarray data. We also applied our method to *Saccharomyces cerevisiae* gene expression data as a real application. We succeeded in estimating more accurate networks than the previous method in both cases.

METHOD

Bayesian network model

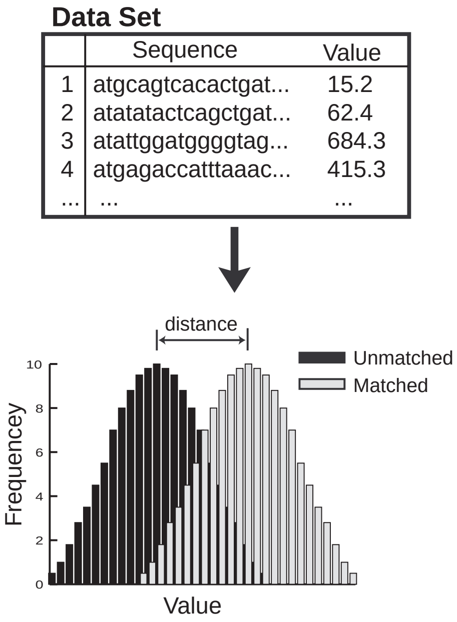
In this subsection, we introduce a Bayesian network and nonparametric regression model (Imoto *et al.*, 2002) as an advance method for estimating gene networks. In the context of Bayesian networks, we consider a directed acyclic graph encoding the Markov assumption. A gene corresponds to a random variable shown as a node and gene regulations are represented by edges in the graph. Suppose that we have n sets of microarrays $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of p genes, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a p dimensional gene expression vector obtained by ith microarray, i.e. x_{ij} is an expression value of jth gene measured by ith microarray. A Bayesian network and nonparametric regression model (Imoto *et al.*, 2002, 2003a) captures the interaction between genes by using nonparametric additive regression models of the form $x_{ij} = m_{j1}(p_{i1}^{(j)}) + \cdots + m_{jq_j}(p_{iq_j}^{(j)}) + \varepsilon_{ij}$, $i = 1, \ldots, n$, where $p_{ik}^{(j)}$ is the expression value of kth parent of the jth gene measured by ith microarray and ε_{ij} depends independently and normally on mean 0 and variance σ_j^2. We construct $m_{jk}(\cdot)$ by B-splines of the form $m_{jk}(p_{ik}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{ik}^{(j)})$, $k = 1, \ldots, q_j$, where $\{b_{1k}^{(j)}(\cdot), \ldots, b_{M_{jk},k}^{(j)}(\cdot)\}$ is a prescribed set of B-splines and $\gamma_{mk}^{(j)}$ are parameters. Hence, a joint density of the all genes can be modeled by

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{\{x_{ij} - \sum_{k=1}^{q_j} m_{jk}(p_{ik}^{(j)})\}^2}{2\sigma_j^2}\right],$$

where $\boldsymbol{\theta}_G$ is a parameter vector. If jth gene has no parent genes, we use a parameter μ_j instead of $\sum_{k=1}^{q_j} m_{jk}(p_{ik}^{(j)})$. Advantages of our model are as follows: Our model can analyze gene expression data as continuous data without extra pretreatment of the data such as discretization, which leads to information loss. Even nonlinear relationships between genes can be extracted.

Motif detection

A motif detection method is used to find a consensus motif from a set of genes which *may* be co-regulated by the same transcription factor. A popular method for determining such a set is to first cluster the genes according to their expression patterns, and then look for common motifs in each of the clusters (e.g. Brazma *et al.* (1998)). In our method, however, we would like to exploit the 'higher level' information extracted by the estimation of the network, represented by the *structure* of the network and likelihood scores (which depend on the network structure) that a certain gene is a parent of another.



Data Set

	Sequence	Value
1	atgcagtcacactgat...	15.2
2	atatatactcagctgat...	62.4
3	atattggatggggtag...	684.3
4	atgagaccatttaaac...	415.3
...	...	...

Fig. 2. Concept of string pattern regression.

Using the structure of the currently estimated network, we choose several genes as putative TFs. We also select sets of genes as candidates genes that may be regulated by each TF, and therefore may contain a common motif. Each of these genes are paired with a score which represents the likelihood that the gene is a direct child of the TF. Details of the likelihood score and the selection process is given in **Criterion** and **Algorithm**.

Since the network structure and the likelihoods that each gene is a direct child of the TF should be *fairly* accurate, we want to find motifs which appear in genes with relatively higher likelihood scores, and does not appear in relatively lower likelihood scores. This kind of motif search is possible using a method called *string pattern regression* proposed in (Bannai *et al.*, 2002), which looks for motifs that separate the set of strings so that the distribution of a numerical attribute paired with the string is best split. More precisely, given a data set $D \subseteq \Sigma^* \times \mathbf{R}$ consisting of pairs of a string and a numerical attribute, the method looks for a pattern p which minimizes the following score:

$$MSE(D, p) = \frac{\sum_{(s,r)\in D_p}(\mu(D_p)-r)^2+\sum_{(s,r)\in D_{\bar{p}}}(\mu(D_{\bar{p}})-r)^2}{|D|},$$

where D_p is the subset of D whose string contains p, $D_{\bar{p}} = D - D_p$, and $\mu(D') = \frac{\sum_{(s,r)\in D'} r}{|D'|}$ is the average of the numeric attributes in any set $D' \subseteq D$.

In this paper, we use the *substring pattern class* as the motif model. A substring pattern is essentially a string of certain length containing no mismatches or ambiguities. The string pattern regression problem can be optimally solved very efficiently for this motif model: in linear time in the total length of the input strings (Bannai *et al.*, 2003).

Although more flexible patterns (e.g. PSSM) are usually preferred, it is known that transcription factor binding sites contain core short patterns which are well conserved with low internal variation (Bussemaker *et al.*, 2001; Keleş *et al.*, 2002). Selection of an appropriate motif model for our method is a difficult problem, and deserves further investigations.

Criterion for choosing a network

In a Bayesian statistical framework, we can choose an optimal gene network based on the posterior probability of a network G

$$\pi(G|\boldsymbol{X}) = \pi(G, \boldsymbol{X})/\pi(\boldsymbol{X}), \qquad (1)$$

where $\boldsymbol{X}$ is the microarray data,

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

Here $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ is a prior distribution on the parameter $\boldsymbol{\theta}_G$ with a hyperparameter vector $\boldsymbol{\lambda}$ and $\mathcal{G}$ is the set of possible networks. Since the network selection does not depend on the denominator of (1), we can use $\pi(G, \boldsymbol{X})$ as a model selector.

The integral in $\pi(G, \boldsymbol{X})$ is the marginal likelihood and represents the fitness of the model to the microarray data. The information of regulatory motifs is stored in $\pi(G)$, which is the prior probability of the network G. Imoto *et al.* (2003b) provided a general framework for combining microarrays and biological knowledge for estimating a gene network based on Bayesian networks. We briefly introduce their method here and show how to incorporate motif information into their framework.

They defined a network energy $E(G) = \sum_{(i,j)\in G} U_{ij}$, where U_{ij} is an energy of an edge from the ith gene to the jth gene, and assume that the probability of the network depends on the Gibbs distribution $\pi(G) = Z^{-1}\exp(-\zeta E(G))$, where $Z = \sum_{g\in\mathcal{G}}\exp(-\zeta E(G))$ is a partition function and ζ a hyperparameter. Under their framework, we can allocate the different energies according to the information of consensus motifs. Concretely, for each ζU_{ij}, we set a value ζ_1 for relationships with motif evidence and ζ_2 otherwise. Note that $0 < \zeta_1 < \zeta_2$. Hence we obtain a prior probability of a network reflecting motif

Step 1. Estimate a gene network from microarray data alone using Bayesian network model.
Step 2. For each gene g, let D_g be the set of child and grand-child genes of g. Genes with $|D_g| \geq 4$ are considered as TFs, and search for motifs in D_g.
Step 3. For each TF, based on the result of the motif detection:

  A) If a parent of the TF contains the motif, we reverse the edge and make it a direct child.
  B) If a grand-child of the TF contains the motif, we add an edge and make it a direct child.

  We also embed this information into Equation (2).

Step 4. Estimate a gene network again along with the motif information.
Step 5. Continue Step 2 through 4 until the network does not change.

Fig. 3. Algorithm for estimating a gene network from microarray data with promoter detection.

information of the form

$$\pi(G) = Z^{-1} \prod_{j=1}^{p} \prod_{i \in L_j} \exp(-\zeta_{\alpha(i,j)}), \qquad (2)$$

where L_j is an index set of parent genes of ith gene and the function $\alpha(i, j)$ takes 1 if jth gene has a motif against ith gene or 2 otherwise. For example, if gene$_2$, gene$_3$ and gene$_4$ have a consensus motif against gene$_1$, but gene$_5$ does not have. We find $\alpha(1, 2) = \alpha(1, 3) = \alpha(1, 4) = 1$ and $\alpha(1, 5) = 2$.

By computing the integral in $\pi(G, \boldsymbol{X})$, we can use it as a network selector. We apply the Laplace approximation to compute this integral and the criterion then results in BNRC (Bayesian network and Nonparametric Regression Criterion) (Imoto *et al.*, 2002) with motif information. The use of Laplace approximation for computing the marginal likelihood has been investigated by (Davison, 1986; Imoto *et al.*, 2002, 2003a; Konishi *et al.*, 2003; Tinerey *et al.*, 1986).

Algorithm

The algorithm of the method is summarized in Figure 3. In Step 1, from microarray data alone, we estimate an initial gene network using a Bayesian network model (Imoto *et al.*, 2002) described in **Bayesian Network Model**. In subsequent steps, we will revise this network using motif information. In Step 2, we select transcription factor candidates. If a gene in the network has many parents and children, we hypothesize that these genes are transcription factors (TFs) that may regulate other genes by binding to consensus motifs in their promoter region of the DNA sequences. In our method, we select as TFs, genes which have more than 4 child or grand-child genes in the estimated network. Note that we do not limit the number of TF candidates in this step. Next, for each selected TF g, we extract a set of genes which may be co-regulated by g and therefore share consensus motifs. Since the network can contain errors concerning direct connections, we define this set as the child and grand-child genes of gene g, denoted by D_g.

Then, we execute the motif detection method described in the previous section for each set D_g. Scores assigned to genes in D_g are calculated as direct children of TF g. After the motif detection, we search from a set of parent genes of the TF g, a motif found in the motif detection method.

In Step 3, based on the result of the motif detection program, we modify the edges of the network as follows.

A) If the motif is found in a parent of the TF, it is possible that this parent is actually a child of the TF. Therefore, we reverse the direction of such edges.

B) If the motif is found in a grand-child of the TF, then it is possible that this gene is a child of the TF. We remove such edges and add direct edges.

After the modification of edges, we remove all edges from the network, except edges modified in the previous step and edges that connect with genes having the motifs. This is done because the greedy hill climbing algorithm used in the Bayesian network estimation method, depends on the initial state of the network before the estimation.

Finally, in Step 4, we estimate the network using the Bayesian network method again, this time along with prior knowledge about the existence of the motif. For the prior probability (Equation (2)), we use ζ_1 for parent-child relations which are supported by the detected motif, and ζ_2 otherwise. Note that, the modifications for the edges do not always remain in the next network estimation. Because the motif detection method does not always succeed in detecting real motifs, we can not blindly trust the result of the detection. Besides, it is possible that a set of genes D_g do not even have any consensus motif. Estimating the network along with a prior information of the motif existence can be considered to be the evaluation of the motif detection using a Bayesian model and microarray expression data.

After the re-estimation of the network, we also execute the motif detection method again. We continue this iteration until the motif detection method does not detect any motif that can affect the result of the next network estimation.

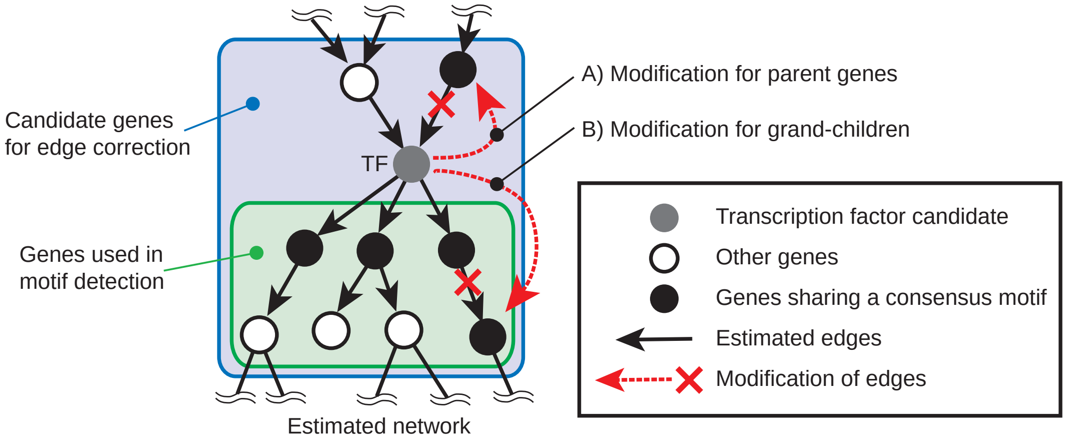
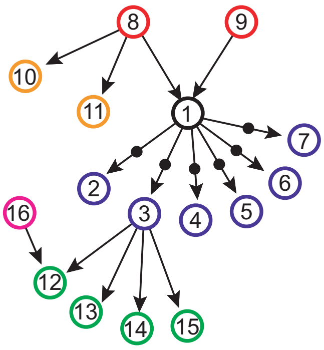


Fig. 4. Brief explanation for modifications of edges. The gray node represents a transcription factor (TF). The motif detection performs to sets of TF's child and grand-child genes (indicated by the green region). Black nodes indicates genes sharing a consensus motifs found in the motif detection. After the motif detection, our method search the motif from parents of the TF. The candidate genes for the edge modification are indicated by the blue region. Red edges represent new edges by the modification.



$$\mathtt{gene3} = \begin{cases} -1 + \varepsilon_3, & \mathtt{gene1} \le -0.5 \\ \mathtt{gene1} + \varepsilon_3, & -0.5 < \mathtt{gene1} \le 0.5 \\ 1 + \varepsilon_3, & 0.5 < \mathtt{gene1} \end{cases}$$

$$\mathtt{gene4} = \begin{cases} 0.4\,\mathtt{gene1} + 1.0 + \varepsilon_4, & \mathtt{gene1} \le -0.3 \\ (\mathtt{gene1} + 1)^2 + \varepsilon_4, & -0.3 < \mathtt{gene1} \le 0.3 \\ 0.4\,\mathtt{gene1} + 1.0 + \varepsilon_4, & 0.3 < \mathtt{gene1} \end{cases}$$

$$\mathtt{gene10} = \begin{cases} 0.4\,\mathtt{gene8} + 1.0 + \varepsilon_{10}, & \mathtt{gene8} \le -0.3 \\ (\mathtt{gene8} + 1)^2 + \varepsilon_{10}, & -0.3 < \mathtt{gene8} \le 0.3 \\ 0.4\,\mathtt{gene8} + 1.0 + \varepsilon_{10}, & 0.3 < \mathtt{gene8} \end{cases}$$

$$\mathtt{gene14} = \begin{cases} 0.4\,\mathtt{gene3} + 1.0 + \varepsilon_{10}, & \mathtt{gene3} \le -0.3 \\ (\mathtt{gene3} + 1)^2 + \varepsilon_{10}, & -0.3 < \mathtt{gene3} \le 0.3 \\ 0.4\,\mathtt{gene3} + 1.0 + \varepsilon_{10}, & 0.3 < \mathtt{gene3} \end{cases}$$

$$\mathtt{gene15} = \begin{cases} 0.2\,\mathtt{gene3} - 1.0 + \varepsilon_{15}, & \mathtt{gene3} \le -0.2 \\ 1.4\,\mathtt{gene3} + \varepsilon_{15}, & -0.2 < \mathtt{gene3} \end{cases}$$

$$\mathtt{gene1} = 1.2\,\mathtt{gene8} + 0.8\,\mathtt{gene9} + \varepsilon_1$$

$$\mathtt{gene2} = 0.6\,\mathtt{gene1} + \varepsilon_2$$

$$\mathtt{gene5} = \cos(1.4\,(\mathtt{gene1} + 3.7)) + \varepsilon_5$$

$$\mathtt{gene6} = 0.6\,\mathtt{gene1} + \varepsilon_6$$

$$\mathtt{gene7} = 0.7\,\mathtt{gene1} + \varepsilon_7$$

$$\mathtt{gene8} = \varepsilon_8, \mathtt{gene9} = \varepsilon_9, \mathtt{gene16} = \varepsilon_{16}$$

$$\mathtt{gene11} = 1.0/(1 + \exp(-4\mathtt{gene8})) + \varepsilon_{11}$$

$$\mathtt{gene12} = 0.8\,\mathtt{gene16} + 0.6(\sin\mathtt{gene3}) + \varepsilon_{12}$$

$$\mathtt{gene13} = 1.3\,\mathtt{gene3} + \varepsilon_{13}$$

Fig. 5. Designed network (left) and its relations assigned to genes (right). Small circles on the edges of the network represent that they share a consensus motif. We assume that `gene1` (node with number 1) is a transcription factor. ε_i in the functions represents noise.

Figure 4 represents an example of the modification of edges. The gray node represents a transcription factor candidate. The motif detection method performs for child and grand-child genes of the TF (genes in the green region). Black nodes indicate that they share a consensus motif. Solid lines are the estimated regulations by a Bayesian network model. In this example, a motif found in TF's children are also found in a TF's parent and in a grand-child. In this case, we reverse the direction of the edge between the TF and the parent, and connect a new direct edge between the TF and the grand-child which has the motif. Dashed red lines represent new edges after such modifications.

Implementation and computational resources

We implemented our program using C++ for the Bayesian network estimation, Objective Caml for the motif detection. The computation was conducted under Sun Fire 15k with 96 CPUs, and Intel Xeon cluster system with 64 CPUs. The program can run parallel on these CPUs using MPI.

COMPUTATIONAL EXPERIMENTS

Monte Carlo simulations

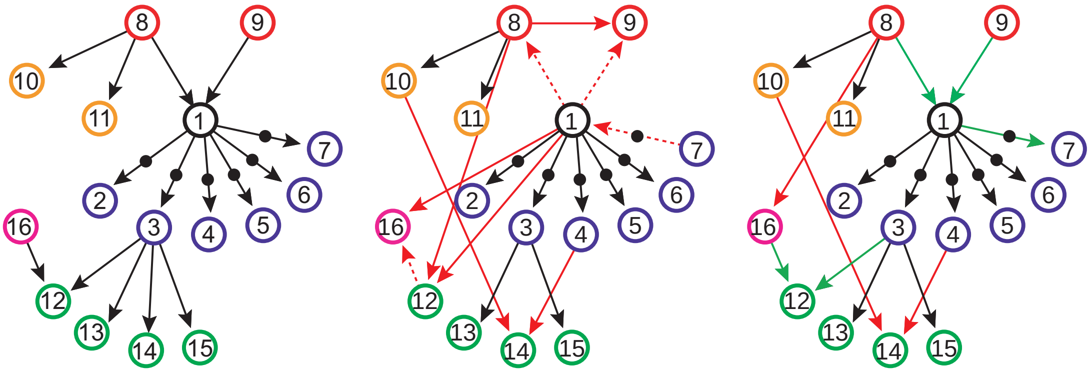
To evaluate the effectiveness of our method, we have conducted Monte Carlo simulations.

Data

We designed an artificial network whose relations of the regulations between genes are shown in Figure 5. The network we designed has 16 genes, and we assume `gene1` (the black node in Fig. 5) to be a transcription factor and their children (the blue nodes) to have a consensus motif. We randomly generate pseudo DNA sequences

Table 1. Performance of the network estimation with or without motif information

	experiments	correct	misdirect	false positive	sensitivity	specificity
(I)	with motif info (1000)	10 768	2086	4 943	71.8 %	54.0 %
(II)	without motif info (1000)	10 639	2898	12,727	70.9 %	38.4 %
(III)	`tatat` detected in (I) (433)	4 785	823	2 118	73.7 %	55.6 %
(IV)	`tatat` not detected in (I) (567)	5 983	1263	2 825	70.3 %	52.8 %



(a) true network (b) without motifs (c) with motifs

Fig. 6. (a) true network, (b) a network estimated without motif information, (c) a network estimated with motif information. Red dashed edges represent mis-directed edges. Red solid edges represent false positive (wrongly estimated edges). Green edges in (c) represent correctly revised edges from (b).

for each gene, and embed a pseudo consensus motif ‘`tatat`’ in `gene2` ~ `gene7` (children of `gene1`) by hand. We eliminate this motif from sequences of other genes. The length of the pseudo DNA sequences is 100 base pairs for all genes. We generated pseudo 100 microarrays for one data set using this network, and we prepared 1000 sets of such data. ε_i in the functions appeared in Figure 5 represents noise for each node. The amount of noise we embedded was set to a signal to noise ratio of 0.3. We ignore motifs whose length is less than 4, since, although motifs of such lengths may represent a biologically significant motif in real organisms, they are most likely a product of chance in our simulation.

For prior probabilities to this Monte Carlo simulation, we use 1.0 for ζ_1, 7.0 for ζ_2. The energies we used were chosen from an experimental viewpoint. When we used a smaller energy (e.g. 2.0) as ζ_2, the motif information could not contribute to the network revision. On the other hand, when we used a larger ζ_2 (e.g. 20.0), the resulting network reflected the motif information too strongly. We observed that our energies, $\zeta_1 = 1.0$ and $\zeta_2 = 7.0$, are not fatalistic, but give appropriate effects for the network revision.

Results

The results of the Monte Carlo simulations are summarized in Table 1. Rows (I) and (II) represent the result of the estimation with or without the motif information. Column ‘specificity’ is the percentage of correctly estimated edges out of the total number of estimated edges, and ‘sensitivity’ is the percentage of correctly estimated edges out of the total number of true edges.

By combining microarray data with the motif information, the specificity increased drastically (38.4 % → 54.0%). Although the number of correct edges only increased slightly (10 639 → 10 768), the number of false positives extremely decreased (12 727 → 4934).

The number of experiments that successfully detected the embedded motif ‘`tatat`’ was 433 times out of 1,000 experiments ((III) in Table 1). When comparing (II) with (IV), we can see that our method could increase the specificity even if the method failed to detect the embedded motif. We observed that for the majority of (IV), our method detected the motif from a subset of `gene1`’s children and therefore an incorrect motif does not lead to serious problems.

```
 1: TF : gene1
 2: Detecting motif from ...
 3:   gene    BNRC score
 4: --------------------
 5:   gene6   152.098
 6:   gene2   194.65
 7:   gene4   231.136
 8:   gene5   124.904
 9:   gene8   227.031
10:   gene9   281.758
11:   gene16  298.498
12:   gene12  254.1
13:   gene3   141.219
14:   gene14  263.637
15:   gene10  272.292
16:   gene11  269.644
17:   gene15  199.679
18:   gene13  196.396
19:
20: Executing the motif
21:       detection method...
22: found motif : tatat
23: matched genes
24:    gene6   gene2   gene4
25:    gene5   gene3
26:
27: Search from parents :
28:   gene7
29: tatat found in gene7
30:
31: Modifying the network...
32: Reverse: gene7 <--> gene1
33: Keep   : gene1 -> gene2
34: Keep   : gene1 -> gene3
35: Keep   : gene1 -> gene4
36: Keep   : gene1 -> gene5
37: Keep   : gene1 -> gene6
38:
39: estimating the next ...
```

Fig. 7. Execution log of the method of the example in Figure 6. Lines from 5 to 18 represent genes and BNRC scores passed to the motif detection method. `tatat` in Line 22 represents the motif found from this gene set. The parent gene of `gene1` in the initial network is only `gene7` in this example. The motif was also found in `gene7` (Line 29). After the motif detection, the method revised the edges based on the existence of the motif (Line 32 $\sim$ 37).

Figure 6 represents a typical result of the Monte Carlo simulations. Figure 6a is the true network we designed, same as in Figure 5b is an initial network estimated by a Bayesian network model using microarray data alone. By extracting the motif information and using a Bayesian network method repeatedly, we obtain a final network shown in Figure 6c. There are four misdirected edges (represented by red dashed arrows) in the network (b), but all of them are revised correctly in (c) (represented in green arrows). Whereas there are 6 falsely estimated edges in (b), after the revision the number of false positives becomes 3, and represented by red edges in (c).

The edge from `gene1` to `gene12` was estimated in the initial network (b) as a direct regulation. This edge was rejected in the re-estimation by a Bayesian network method, and a correct edge from `gene3` to `gene12` was added. The correction for the direction of the edge from `gene7` to `gene1` in (b) results in the correction for regulation between `gene1` and its parents. This correction also revises the indirect relation from `gene1` to `gene12` via `gene8`. The log of the execution of our method is represented in Figure 7.

Application to real data

Data

We applied our method to *Saccharomyces cerevisiae* microarray data obtained by disrupting 100 genes, most of which are transcription factors (Imoto *et al.*, 2003a). We focused on three transcription factors, *CHA4*, *GAL11*, and *SWI6*, that have many child genes in the estimated network of Imoto *et al.* (2002), because these genes probably play important roles in the gene regulations. We extracted 124 genes that have distance less than or equal to two from the above three genes. The promoter region of their DNA sequences are retrieved from GenBank database.

Results

Our method repeated the network estimation and motif detection four times with this data. Since *CHA4* was selected as a TF for all iterations, we focus on *CHA4* to evaluate our method. Figures 8 and 9 show the partial network in the neighborhood of *CHA4*, estimated by the Bayesian network model without, and with the motif information, respectively. In both figures, the function of each gene is indicated by a 2 digit number, which corresponds to the MIPS functional category (Mewes *et al.*, 2002). For example *TOP2* located on the right side of Figure 9 has the function ‘cell cycle and DNA processing’ and ‘subcellular localization’.

In the four iterations, our algorithm detected the motifs: `aaaga`, `aaacg` (twice), and `taaac`. Surprisingly, the last motif is known as a promoter element of an yeast cell cycle transcription factor SFF (Swi Five Factor) (Pic *et al.*, 2000). Black nodes in Figures 8 and 9 indicate that they have the consensus motifs `taaac`. *ACE2* is a gene known to be regulated by SFF and contains SFF promoter elements (Pic *et al.*, 2000). Though this gene was not selected as a gene set for the motif detection, it became a child of *RIM11* in the revised network.

CHA4, which is selected as a TF candidate, has functions of the cell cycle and metabolism. Most of the genes located downstream of *GAL11* and *CHA4* in both networks have functions related to the cell cycle and metabolism.

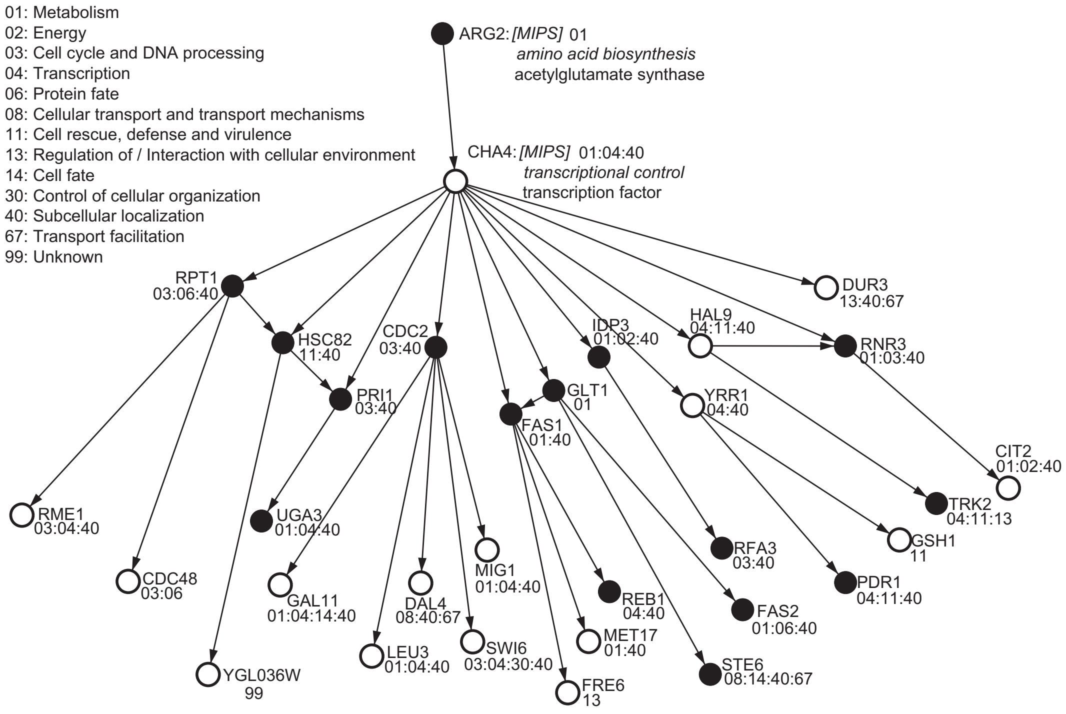


Fig. 8. A partial network estimated using microarray data alone.

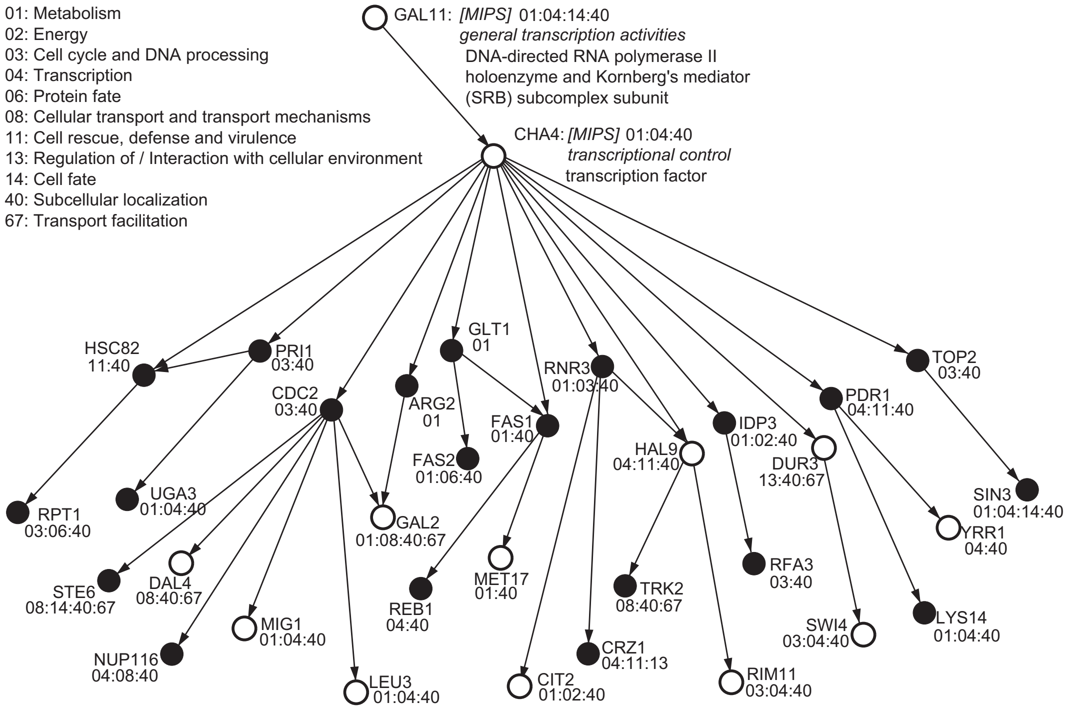


Fig. 9. A partial network estimated using microarray data and motif information.

Table 2. Alignment of the detected motif with known genes. Capital letters are consistent with the known consensus motifs

	MCM1 SFF
motif	`CCY-WWWNN-RG RYMAAYA`
ACE2	`CtC-AAAA-CGGcaaaat-GTAAACAttggc`
HOF1	`tCC-TcTT-TGGgcaagttGTAAACAataaa`
ALK1	`CCC-TTTT-TGGtaaaa-cGTAAACAaaata`
SUR7	`CCC-AATCG-GGaaaa-ttGTAAACAttagc`
BUD4	`CCC-gATTT-GGaaaaa-gGTAAACAacaat`
SWI5	`CCT-gTTTA-GGaaaaa-gGTAAACAataac`
CLB2	`CC-GAATCA-GGaaaa--gGTCAACAacgaa`
REB1	`CCaaccTAA-AGtaaataaATAAACAtcatc`
ARG2	`CCagTTccACGGcaactcacTAAACctatcc`

Y = C or T, W = A or T, R = A or G, M = A or C

According to the above analysis, although there is no biological evidence that *CHA4* is related to SFF, most black genes have functions related to the cell cycle or metabolism. *CHA4* is also a transcription factor that functions as a cell cycle regulator. We can say that there may be a relation between *CHA4* and SFF.

The *MCM1*-SFF complex regulates the G2 phase of the cell cycle, and *ACE2* is known to have the *MCM1* promoter element, as well as the SFF element (Pic *et al.*, 2000). For all genes which contains the motif `taaac`, we looked for genes which have an *MCM1* binding site near the SFF binding site as in *ACE2*. The result is shown in Table 2. The upper 7 rows are the binding sequences shown in Pic *et al.* (2000). The lower two rows show the genes which exhibit a putative *MCM1* binding site. We can see that the motifs of these two genes are very similar to known *MCM1*-SFF binding sites. Unfortunately, transcription factors *MCM1* and SFF (primary component *FKH2* (Boros *et al.*, 2003)), and genes which they regulate, such as *HOF1*, are not contained in our data set. The estimation of a network for all genes is unrealistic from a statistical point of view, and selection of genes is a very difficult and important problem.

GAL11 is known as a general transcription factor, and is known to regulate *GAL2* (Suzuki *et al.*, 1988). However, in the network estimated without the motif information, *GAL2* lies upstream of *GAL11*, as a parent of *ARG2* (data not shown). Interestingly, in the revised network, *GAL11* moved to an upstream location of the network, compared to that of the network without motif information, and we can see that the relation between *GAL11* and *GAL2* is corrected.

CONCLUSION

We proposed a statistical method for estimating gene networks, combining microarray gene expression data and DNA sequences of regulatory regions of genes. From the Monte Carlo simulations, we can conclude that our method can estimate more accurate networks than existing methods, and can simultaneously detect the promoter elements. We observed that the motif information is useful for revising some incorrect relations in the network estimated by microarray data alone. In a real data application, we succeeded in estimating a gene network which contains known regulatory relations, and we could detect a known motif as well. We also observed in both Monte Carlo simulations and real data experiments, that the effect of small corrections made based on the motif information seemed to propagate through the entire network, rather than modify a local neighborhood of where the motif was detected.

Our method also has an advantage as a motif detection method. Determining the set of co-regulated genes that may have a consensus motif is a difficult problem, because indirectly regulated genes may be included and/or directly regulated genes may be excluded (Holmes and Bruno, 2000; Bussemaker *et al.*, 2001). Using a Bayesian network model, we can roughly determine the direct/indirect relation between genes. Therefore, our method is another approach for solving this problem to obtain more biologically meaningful results.

There are several works combining gene expression profiles with promoter element information to investigate gene networks. In Segal *et al.* (2002), a probabilistic framework was proposed, that models the process by which transcriptional binding explains the expression of genes. In Pilpel *et al.* (2001), they show a strategy to find motif combinations which effect the gene expression. In Hartemink *et al.* (2002), data from genomic location analysis is combined in the inference of the network. Our method is different from these methods, and the uniqueness of our method lies in the interactive improvement of Bayesian network and promoter element detection.

From a biological point of view, the actual machinery of the regulation in the organism is more complicated. For example, transcription factors are often realized by a complex consisting of a set of proteins. Our Bayesian network model cannot treat protein complexes. We would like to investigate this topic in our future research.

REFERENCES

Akutsu,T., Kuhara,S., Maruyama,O. and Miyano,S. (2003) Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. *Theor. Comput. Sci.*, **298**, 235–251. (Preliminary version has appeared in *Proc. 9th ACM-SIAM Symp. Discrete Algorithms*, **56**, 695–702, 1998)

Akutsu,T., Miyano,S. and Kuhara,S. (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac. Symp. Biocomput.*, **4**, 17–28.

Akutsu,T., Miyano,S. and Kuhara,S. (2000a) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J. Comp. Biol.*, **7**, 331–344.

Akutsu,T., Miyano,S. and Kuhara,S. (2000b) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.

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

Bannai,H., Inenaga,S., Shinohara,A., Takeda,M. and Miyano,S. (2003) Efficiently finding regulatory elements using correlation with gene expression. submitted for publication

Boros,J., Lim,F.L., Darieva,Z., Pic-Taylor,A., Harman,R., Morgan,B.A. and Sharrocks,A.D. (2003) Molecular determinants of the cell-cycle regulated Mcm1p-Fkh2p transcription factor complex. *Nucleic Acids Res.*, **31**, 2279–2288.

Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.*, **8**, 1202–1215.

Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Chen,T., He,H.L. and Church,G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **4**, 29–40.

Cooper,G. and Herskovits,E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.

Davison,A.C. (1986) Approximate predictive likelihood. *Biometrika*, **73**, 323–332.

De Hoon,M.J.L., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2003) Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pac. Symp. Biocomput.*, **8**, 17–28.

Friedman,N. and Goldszmidt,M. (1998) *Learning Bayesian Networks with Local Structure*. Jordan,M.I. (ed.), Kluwer Academic Publisher, pp. 421–459.

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

Hartemink,A.J., Gifford,D.K., Jaakkola,T.S. and Young,R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, **6**, 422–433.

Hartemink,A.J., Gifford,D.K., Jaakkola,T.S. and Young,R.A. (2002) Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.*, **7**, 437–449.

Heckerman,D. (1998) *A Tutorial on Learning with Bayesian Networks*. Jordan,M.I. (ed.), Kluwer Academic Publisher, pp. 301–354.

Holmes,I. and Bruno,W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. In *Proceedings of Intelligent Systems for Molecular Biology*. pp. 202–210.

Imoto,S., Goto,T. and Miyano,S. (2002) Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. *Pac. Symp. Biocomput.*, **7**, 175–186.

Imoto,S., Kim,S., Goto,T., Aburatani,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003a) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinform. Comput. Biol.*, in press. (Preliminary version has appeared in *Proc. 1st IEEE Computer Society Bioinformatics Conference*, 219–227, 2002)

Imoto,S., Higuchi,T., Goto,T. and Miyano,S. (2003b) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. In *Proceedings of 2nd IEEE Computer Society Bioinformatics Conference*. in press.

Keleş,S., van der Laan,M. and Eisen,M.B. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.

Kim,S., Imoto,S. and Miyano,S. (2003) Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. In *Proceedings of the 1st International Workshop on Computational Methods in Systems Biology*, Lecture Note in Computer Science, 2602, Springer, pp. 104–113.

Konishi,S., Ando,T. and Imoto,S. (2003) Bayesian information criteria and smoothing parameter selection in radial basis function networks. submitted for publication

Maki,Y., Tominaga,D., Okamoto,M., Watanabe,S. and Eguchi,Y. (2001) Development of a system for the inference of large scale genetic networks. *Pac. Symp. Biocomput.*, **6**, 446–458.

Mewes,H.W., Frishman,D., Güldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Münsterkoetter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

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

Pic,A., Lim,F.L., Ross,S.J., Veal,E.A., Johnson,A.L., Sultan,M.R.A., West,A.G., Johnston,L.H., Sharrocks,A.D. and Morgan,B.A. (2000) The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *EMBO*, **19**, 3750–3761.

Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Segal,E., Barash,Y., Simon,I., Friedman,N. and Koller,D. (2002) From promoter sequence to expression: a probabilistic framework. In *Proceedings of the International Conference on Research in Computational Molecular Biology '02*. pp. 273–280.

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

Suzuki,Y., Nogi,Y., Abe,A. and Fukasawa,T. (1988) GAL11 protein, an auxiliary transcription activator for genes encoding galactose-metabolizing enzymes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **8**, 4991–4999.

Tinerey,L. and Kadane,J.B. (1986) Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.*, **81**, 82–86.