Journal of Bioinformatics and Computational Biology
Vol. 3, No. 6 (2005) 1295–1313

Imperial College Press
www.icpress.co.uk

UTILIZING EVOLUTIONARY INFORMATION AND GENE EXPRESSION DATA FOR ESTIMATING GENE NETWORKS WITH BAYESIAN NETWORK MODELS

YOSHINORI TAMADA[*,‡], HIDEO BANNAI[†,§], SEIYA IMOTO[†,¶], TOSHIAKI KATAYAMA[†,‖], MINORU KANEHISA[*,**] and SATORU MIYANO[†,††]

[*]*Bioinformatics Center, Institute for Chemical Research*
Kyoto University, Gokasho, Uji
Kyoto, 611-0011, Japan

[†]*Human Genome Center, Institute of Medical Science*
The University of Tokyo, 4-6-1, Shirokanedai
Minato-ku, Tokyo, 108-8639, Japan
[‡]*tamada@kuicr.kyoto-u.ac.jp*
[§]*bannai@ims.u-tokyo.ac.jp*
[¶]*imoto@ims.u-tokyo.ac.jp*
[‖]*ktym@hgc.jp*
[**]*kanehisa@kuicr.kyoto-u.ac.jp*
[††]*miyano@ims.u-tokyo.ac.jp*

Received 23 February 2005
Revised 29 June 2005
Accepted 29 June 2005

Since microarray gene expression data do not contain sufficient information for estimating accurate gene networks, other biological information has been considered to improve the estimated networks. Recent studies have revealed that highly conserved proteins that exhibit similar expression patterns in different organisms, have almost the same function in each organism. Such conserved proteins are also known to play similar roles in terms of the regulation of genes. Therefore, this evolutionary information can be used to refine regulatory relationships among genes, which are estimated from gene expression data. We propose a statistical method for estimating gene networks from gene expression data by utilizing evolutionarily conserved relationships between genes. Our method simultaneously estimates two gene networks of two distinct organisms, with a Bayesian network model utilizing the evolutionary information so that gene expression data of one organism helps to estimate the gene network of the other. We show the effectiveness of the method through the analysis on *Saccharomyces cerevisiae* and *Homo sapiens* cell cycle gene expression data. Our method was successful in estimating gene networks that capture many known relationships as well as several unknown relationships which are likely to be novel. Supplementary information is available at http://bonsai.ims.u-tokyo.ac.jp/~tamada/bayesnet/.

Keywords: Gene network; evolutionary information; Bayesian network; microarray data.

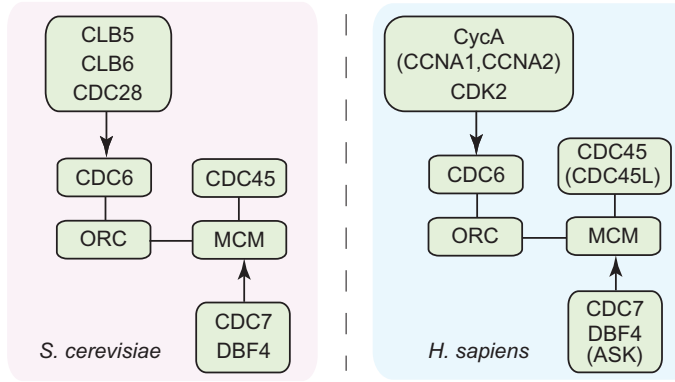
[*]Corresponding author.

1. Introduction

Computational inference of gene networks from microarray data has received considerable attention in the literature,[1] and it has been applied to drug target gene discovery.[2,3] Methods for automatically constructing such networks from genomic data are invaluable tools for understanding the mechanisms of organisms' cells. A Bayesian network is a statistical model which has been successfully applied to the estimation of gene networks from gene expression data.[4–9] However, since information obtained from microarrays is limited and a large number of parameters contained in their models must be estimated from a relatively small number of microarrays, it is difficult to obtain biologically accurate gene networks. A possible solution for this problem is to combine gene expression data with other biological information. Several methods have been proposed for this direction. e.g. location information,[10] use of biological knowledge as a prior in Bayesian network models,[2] protein-protein interaction,[11] promoter sequences.[12] Although these methods have given successful results, they have only been applied to simple organisms such as *S. cerevisiae* and bacteria, because they require, in addition to microarray data, comprehensive data sets which are organism specific and are not readily available for higher organisms. Therefore, developing methods for incorporating information sources closer at hand is a very important problem.

Sequences produced and made available by various genomic projects are representative of such information, and are becoming widely used in comparative genomics. Furthermore, recent studies have revealed that orthologous genes which show similar expression patterns have strongly related functions in each of the organisms' cells.[13–16] These studies indicate that important core regulatory relationships of genes which are required for living cells may also be conserved even in diverse organisms. For example in the cell cycle pathways for *S. cerevisiae* and *H. sapiens* in the KEGG database,[17] we observe that many edges are common between orthologous gene pairs in the two organisms (Fig. 1). Thus, there arises a problem of how to utilize such evolutionarily conserved information in order to estimate gene networks from gene expression data more accurately.

This paper proposes a statistical method that estimates gene networks from gene expression data and the information of evolutionarily conserved proteins among distinct organisms using Bayesian network models. Our method utilizes the evolutionary information between two organisms and simultaneously estimates two gene networks so that the network of one organism helps to estimate the other organism's network and *vice versa*. In order to show the effectiveness of our method, we focus on well investigated and publicly available datasets, and apply our method to estimate the cell cycle gene networks of *S. cerevisiae* and *H. sapiens*.[18,19] Using the cell cycle time-course gene expression data of these two species, the dynamic Bayesian network model with nonparametric regression[9] is employed as the network model. Since *H. sapiens* is a much more complicated organism than *S. cerevisiae*, it is more difficult to estimate regulatory relations between genes of *H. sapiens* than those of



S. cerevisiae	*H. sapiens*	BLAST E-value
CLB5	CCNA1	4.03×10^{-25}
CLB6	CCNA2	2.73×10^{-25}
CDC28	CDK2	1.65×10^{-85}
CDC6	CDC6	2.31×10^{-32}
CDC45	CDC45L	1.91×10^{-32}
CDC7	CDC7	4.35×10^{-16}
DBF4	ASK	3.10×10^{-6}
MCM	MCM	$2.00 \times 10^{-156*}$
ORC	ORC	$4.70 \times 10^{-11*}$

*E-values of MCM and ORC are the largest E-values of genes in the complexes.

Fig. 1. An example of conserved relationships between *S. cerevisiae* and *H. sapiens*.

S. cerevisiae from information on gene expression data alone. Through the comparison using the KEGG cell cycle pathways and annotations from the gene ontology hierarchy,[20] we have confirmed that our method succeeded in estimating these gene networks more accurately than the previous method. Moreover, our method also succeeded in estimating highly possible and unknown relationships which are likely to be novel.

2. Method

The purpose of this section is to define a statistical framework for utilizing microarray data and evolutionary information for estimating networks of two distinct organisms under Bayesian statistics.

2.1. *Bayesian network model*

A gene network, or a gene regulatory network, is a graphical model that represents the regulatory relationships between genes. In a gene network, if there is an edge

from gene a to gene b, then the edge represents that gene a regulates gene b, or the expression of gene b depends on the expression of gene a. We model a gene network G as a Bayesian network, where genes are represented by random variables and the structure is described as a directed graph with the random variables as its nodes. Let $\mathcal{X} = \{X_1, X_2, \ldots, X_p\}$ be a set of random variables (genes) in the network G, where p is the number of nodes. In the context of Bayesian networks, the joint probability of $\mathcal{X}$ conditional on G can be decomposed as a product of conditional probabilities,

$$P(\mathcal{X}|G) = \prod_{j=1}^{p} P(X_j|Pa(X_j)), \tag{1}$$

where $Pa(X_j)$ is a set of parent variables of X_j in G. Suppose that $\boldsymbol{X}$ is a gene expression data matrix whose element x_{ij} corresponds to the expression value of the jth gene in the ith array, where $i = 1, \ldots, n$, $j = 1, \ldots, p$. Here, n and p represent the number of microarrays and genes, respectively. Since microarray data take continuous variables, the probabilistic measures in Eq. (1) are replaced by densities and the likelihood of $\boldsymbol{X}$ conditional on G is given by

$$P(\boldsymbol{X}|G) = \prod_{i=1}^{n} \prod_{j=1}^{p} p(x_{ij}|pa(x_{ij})), \tag{2}$$

where $pa(x_{ij})$ is a set of expression values of the parent genes of jth gene at ith experiment.

2.2. *Probabilistic framework*

We consider two organisms A and B and focus on a set S_A of genes of A and a set S_B for B. The evolutionary information $\boldsymbol{H}_{AB}$ between A and B is given as the set of gene pairs of A and B defined by

$$\boldsymbol{H}_{AB} = \{(a, b) \mid \mathrm{e}(a, b) < \delta,\ a \in S_A,\ b \in S_B\}, \tag{3}$$

where $\mathrm{e}(a, b)$ represents the E-value between gene a and gene b given by the BLAST[21] search, and δ a threshold. We consider that the gene pairs included in $\boldsymbol{H}_{AB}$ are orthologous gene pairs among organisms A and B. In the BLAST search, we will obtain different E-values in the two cases that we search gene a from organism B, and gene b from organism A, and we take $\mathrm{e}(a, b)$ as the larger E-value. For defining $\boldsymbol{H}_{AB}$, we have also tried to use predefined orthologous gene pair lists available in databases such as NCBI and KEGG, instead of using the BLAST E-values as described above. However, we did not obtain better results with such orthologous gene pairs. Other methods such as a protein domain based method used in Babu and Teichmann[22] can also be used and could be an interesting problem to investigate, but is out of the scope of this paper.

Assume that we are given $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ which are matrices representing gene expression data for organisms A and B, respectively. Suppose that we want to

estimate two gene networks, G_A and G_B of organisms A and B, respectively. Under the framework of Bayesian statistics, we estimate two networks simultaneously by maximizing the following posterior probability function,

$$P(G_A, G_B | \boldsymbol{X}_A, \boldsymbol{X}_B, \boldsymbol{H}_{AB}). \tag{4}$$

Supposing that $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are independent, the posterior probability Eq. (4) can then be decomposed as

$$\begin{aligned} &P(G_A, G_B | \boldsymbol{X}_A, \boldsymbol{X}_B, \boldsymbol{H}_{AB}) \\ &\quad \propto P(\boldsymbol{X}_A, \boldsymbol{X}_B | G_A, G_B, \boldsymbol{H}_{AB}) P(G_A, G_B, \boldsymbol{H}_{AB}) \\ &\quad = P(\boldsymbol{X}_A | G_A) P(\boldsymbol{X}_B | G_B) P(\boldsymbol{H}_{AB} | G_A, G_B) P(G_A, G_B) \\ &\quad = P(\boldsymbol{X}_A | G_A) P(\boldsymbol{X}_B | G_B) P(\boldsymbol{H}_{AB} | G_A, G_B) P(G_A) P(G_B). \end{aligned} \tag{5}$$

Here $P(\boldsymbol{X}_A | G_A, G_B, \boldsymbol{H}_{AB}) = P(\boldsymbol{X}_A | G_A)$ and $P(\boldsymbol{X}_B | G_A, G_B, \boldsymbol{H}_{AB}) = P(\boldsymbol{X}_B | G_B)$ hold under our model. The first two probabilities in Eq. (5) measure how much the microarray data $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ fit the estimated gene networks G_A and G_B, and are independently calculated by the Bayesian network models. The second probability $P(\boldsymbol{H}_{AB} | G_A, G_B)$ is a posterior probability of $\boldsymbol{H}_{AB}$ given the networks G_A and G_B. The last $P(G_A, G_B) = P(G_A)P(G_B)$ is a prior probability calculated based on the prior knowledge for the networks G_A and G_B. The probability $P(\boldsymbol{H}_{AB} | G_A, G_B)$ can be considered as a likelihood of the evolutionary information $\boldsymbol{H}_{AB}$ given the two networks G_A and G_B. For the prior probabilities $P(G_A)$ and $P(G_B)$, we can use several types of prior information on networks G_A and G_B. If we do not have any information on the networks, we use constant probabilities for $P(G_A)$ and/or $P(G_B)$.[2] In this paper, since we do not assume any structure on the networks, we use constant prior probabilities for G_A and G_B. We next discuss how we construct the probability $P(\boldsymbol{H}_{AB} | G_A, G_B)$.

2.3. *The probability of the evolutionary information given two networks*

The probability $P(\boldsymbol{H}_{AB} | G_A, G_B)$ represents a likelihood of the evolutionary information $\boldsymbol{H}_{AB}$ given two networks G_A and G_B. Suppose that two pairs of genes (a, a'), (b, b') are included in $\boldsymbol{H}_{AB}$. It is expected that the relationship between a and b is the same as the relationship between a' and b'. That is, if gene a regulates gene b in organism A, we may expect that gene a' also regulates gene b' in organism B. On the other hand, if there is no relationship between genes a and b, we expect that there may also be no relationship between a' and b'.

According to this assumption, we represent the likelihood of the evolutionary information $\boldsymbol{H}_{AB}$ given two networks by (I) the number of edges existing in common between orthologous gene pairs (e.g. (i) in Fig. 2), and (II) the number of gene pairs not connected in common between orthologous gene pairs (e.g. (ii) in Fig. 2). We denote these numbers by n_{P} and n_{N} respectively. For example, in Fig. 1, CCNA1 $\rightarrow$ CDC6 of *H. sapiens* and CLB5 $\rightarrow$ CDC6 of *S. cerevisiae* are counted as n_{P}. On the

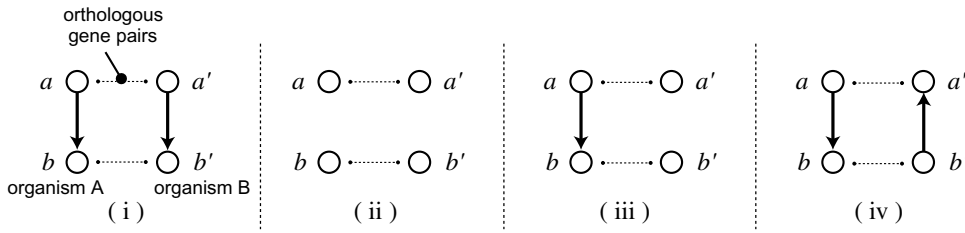


Fig. 2. Examples of edge variations connected between orthologous gene pairs. a and b are genes of organism A, and a' and b' are genes of organism B, where $(a, a'), (b, b') \in \boldsymbol{H}_{AB}$. The relationships of the orthologous gene pairs in (i) and (ii) are conserved. Edges in (iii) and (iv) are inconsistent with each other between these pairs. Therefore, (i) is counted as n_{P}, (ii) is counted as n_{N} in Eq. (6).

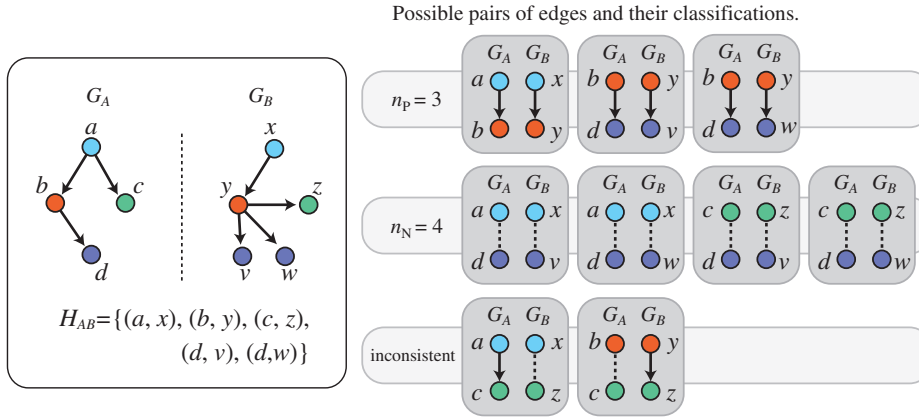


Fig. 3. An example of how we compute n_{P} and n_{N}. Suppose that two networks G_A and G_B, and the evolutionary information $H_{AB} = \{(a, x), (b, y), (c, z), (d, v), (d, w)\}$ are given (left). All possible pairs of edges in G_A and G_B are classified into (i) those connected in common (n_{P}), (ii) those not connected in common (n_{N}), and (iii) those that are inconsistent (right). The dashed lines indicate that they are not connected in the network. In total, $n_{\mathrm{P}} = 3$, $n_{\mathrm{N}} = 4$, and there are two inconsistent pairs in this example.

other hand, genes CCNA1 and CDC45L of *H. sapiens*, and CLB5 and CDC45 of *S. cerevisiae* are not connected in common, and counted as n_{N}. Figure 3 shows an example that explains how we compute n_{N} and n_{P} for a given pair of networks. By using n_{P} and n_{N}, we construct $P(\boldsymbol{H}_{AB}|G_A, G_B)$ as

$$P(\boldsymbol{H}_{AB}|G_A, G_B) = Z^{-1} \exp(\zeta_{\mathrm{P}} n_{\mathrm{P}} + \zeta_{\mathrm{N}} n_{\mathrm{N}}) \,, \tag{6}$$

where Z is the normalizing constant, and ζ_{P} and $\zeta_{\mathrm{N}} (> 0)$ are hyperparameters. These hyperparameters control the balance between gene expression data and evolutionary information.

2.4. *Algorithm*

By using $P(\boldsymbol{X}_A|G_A)$, $P(\boldsymbol{X}_B|G_B)$, $P(\boldsymbol{H}_{AB}|G_A, G_B)$, $P(G_A)$ and $P(G_B)$ described above, we define a criterion $\mathrm{GNS}(G_A, G_B)$ (Gene Network Score) for evaluating gene networks G_A and G_B by taking the logarithm of the joint probability Eq. (4). Precisely, the criterion is defined as

$$\mathrm{GNS}(G_A, G_B) = \log P(\boldsymbol{X}_A|G_A)P(\boldsymbol{X}_B|G_B)P(\boldsymbol{H}_{AB}|G_A, G_B)P(G_A)P(G_B). \quad (7)$$

We estimate the optimal structures of the networks G_A and G_B by maximizing $\mathrm{GNS}(G_A, G_B)$. Actually, finding the optimal pair of networks that maximize $\mathrm{GNS}(G_A, G_B)$ is a very difficult problem, because we need to consider all possible combinations of two networks G_A and G_B for the criterion. In this paper, therefore, we developed a greedy hill-climbing algorithm to find G_A and G_B. The algorithm can be described as follows:

Step 1: We separately estimate G_A and G_B from gene expression data alone by the Bayesian network model, e.g. Imoto *et al.*[7] These networks are used as the initial networks for the following steps.

Step 2: We select randomly a gene from genes of G_A and G_B. We denote the selected gene by a for convenience.

Step 3: Let gene b be a gene of the same organism as gene a. Gene b is a candidate parent of gene a. We consider the following operations between gene a and gene b.

(a) If there does not exist edge $b \to a$ in the network, add gene b as a parent of gene a.

(b) If there exists edge $b \to a$ in the network, remove gene b from parents of gene a.

We test these operations for all possible parent genes b, and perform one operation that gives the largest $\mathrm{GNS}(G_A, G_B)$.

Step 4: Repeat Steps 2 and 3, until any operation does not improve $\mathrm{GNS}(G_A, G_B)$.

Step 5: Repeat from Steps 1 to 4 for the specified times. We obtain a pair of optimal networks G_A and G_B that gives the largest $\mathrm{GNS}(G_A, G_B)$ out of all candidates in Step 4.

For a single gene network with a small number of genes, an efficient algorithm to obtain the optimal network has been proposed,[23,24] and it can be used for Step 1 of our algorithm to obtain the initial networks. In general, however, calculating the optimal pair of networks is infeasible and is an open problem.

3. Computational Experiments

For evaluating the effectiveness of the proposed method, we analyzed *S. cerevisiae* and *H. sapiens* cell cycle microarray data. We conducted two computational experiments with different datasets: the first experiment consists of a small number

of genes that include well-known relationships. The second consists of genes with unknown relationships in addition to the former dataset. The latter experiment is designed for trying to discover unknown relationships in the dataset rather than for evaluating the proposed method. In these experiments, we used publicly available microarray data for *S. cerevisiae*[18] and *H. sapiens*.[19] Both microarray data are cell cycle related and measured as time series data, consisting of 77 time points for *S. cerevisiae*, and 114 time points for *H. sapiens*.

For a gene network model, since we use time series cell cycle gene expression data, we employ the dynamic Bayesian network model with nonparametric regression[9] in this computational experiment, instead of a Bayesian network model. The dynamic Bayesian network can model time dependencies between genes, and is therefore more suitable than the Bayesian network model for using time series data. Let $\boldsymbol{X}$ be a matrix whose (i, j)th element x_{ij} corresponds to the expression value of the jth gene at time i for an organism, where $i = 1, \ldots, T$, $j = 1, \ldots, p$. Here, T and p represent the numbers of microarrays (time points) and genes, respectively. In the context of the dynamic Bayesian networks, the likelihood of $\boldsymbol{X}$ conditional on the network structure G can be decomposed as a product of conditional densities,

$$P(\boldsymbol{X}|G) = \pi(G) \int \prod_{i=1}^{T} \prod_{j=1}^{p} p_j(x_{ij}|pa(x_{i-1,j}), \theta_j)\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})\, d\boldsymbol{\theta}, \tag{8}$$

where $\pi(G)$ is the prior probability of the network G, $pa(x_{0j}) = \emptyset$, $p_j(x_{ij}|pa(x_{ij}), \theta_j)$ the conditional densities for the jth gene, and $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$ the prior distribution on the parameter $\boldsymbol{\theta}$ specified by the hyperparameter $\boldsymbol{\lambda}$.

3.1. *Experiment I: KEGG cell cycle regulated G_1/S phase genes*

Dataset: In this experiment, we focused on G_1/S phase specific genes listed on the KEGG cell cycle pathways. Since both cell cycle pathways of *S. cerevisiae* and *H. sapiens* in the G_1/S phase are very similar, it is relatively easy to evaluate the effectiveness of the method. Below, we refer to these cell cycle pathways in the KEGG database as KEGG pathways. Since it is difficult to estimate relationships between genes whose expression values do not change significantly, we focused on genes that are chosen as "cell cycle-regulated" in Spellman *et al.*[18] and Whitfield *et al.*[19] In addition to these genes, we selected genes that are not listed as cell cycle-regulated but are known to interact with the selected cell cycle-regulated genes from the KEGG pathways. In summary, we collected 17 genes for *S. cerevisiae* and 19 genes for *H. sapiens*, listed in Table 1. According to the previous work,[13] we basically used 10^{-5} for the E-value threshold δ. Since the number of genes is very small, we added a restriction for the construction of evolutionary information $\boldsymbol{H}_{AB}$. If more than four genes show E-values lower than δ for a gene, we employ only up to the fourth gene as the orthologous genes.

Result: First, we evaluated how much the proposed method improves the estimated networks compared to the previous method, which is the dynamic Bayesian

Table 1. List of genes in Experiment I.

S. cerevisiae (17 genes)	*H. sapiens* (19 genes)
CDC20, CDC28, CDC45, CDC46, CDC47, CDC54, CDC6, CDC7, CLB5, CLB6, CLN1, CLN2, MCM2, MCM3, MCM6, ORC1, SIC1	CCNA1, CCNA2, CCND1, CCNE1, CCNE2, CDC25A, CDC45L, CDC6, CDC7, CDK7, CDKN1B, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, ORC1L, ORC3L

network model based on only microarray data. Because it is relatively easy to search the optimal network when using a dynamic Bayesian model instead of a Bayesian network model, and the number of genes for the estimation in this experiment is sufficiently small, we estimated the optimal networks as the initial networks instead of using the greedy algorithm, when we estimate single networks in Step 1 of our algorithm. Therefore, we compared networks estimated by our proposed method with these optimal networks. In the evaluation of the comparison, we focused on how many known relationships kept in the KEGG pathways are estimated. Note that there are many regulatory relationships between genes which do not appear in the KEGG pathways. The hyperparameters ζ_{P} and ζ_{N} were selected such that the proposed method estimates the most consistent gene networks with the KEGG pathways ($\zeta_{\mathrm{P}} = 1.09$, $\zeta_{\mathrm{N}} = 0.31$).

The results of the comparison are summarized in Table 2. The specificity in the table represents the ratio of how many estimated edges are consistent with the KEGG pathways, defined as (specificity) = (# of estimated known edges) / (# of estimated edges). The sensitivity in the table represents the ratio of how many edges in the KEGG pathways are estimated correctly by the method, defined as (sensitivity) = (# of estimated known edges) / (# of known edges). By utilizing the evolutionary information, although the *S. cerevisiae*'s network was not improved very much, we observe that the proposed method improves drastically the sensitivity of *H. sapiens* (0.244 → 0.478), as well as the specificity (0.440 → 0.571). This is because the gene networks of *S. cerevisiae* could be estimated sufficiently from microarray data alone, while *H. sapiens* is much more difficult to estimate from microarray data and could be improved by the

Table 2. Results of Experiment I. "Proposed" represents the results by the proposed method. "Previous" represents results by the previous method.

	Proposed	Previous
S. cerevisiae		
Specificity	**0.500**	0.482
Sensitivity	**0.682**	0.540
H. sapiens		
Specificity	**0.571**	0.440
Sensitivity	**0.478**	0.244

proposed method. From this result, we have confirmed that the information of *S. cerevisiae*'s network was used to improve the estimation of *H. sapiens*'s network.

The estimated networks are shown in Fig. 4. Note that even if we estimate edges not appearing in the KEGG pathways, this does not mean the edges are wrongly estimated. Since the edges existing only in the network estimated by the proposed method are newly estimated edges using the evolutionary information (Table 3), we analyze some of these edges below (Fig. 5). The complete lists of the edges can be found in the supplementary information.

(a) *S. cerevisiae*: CLB6 → CDC6 and *H. sapiens*: CCNA1 → CDC6

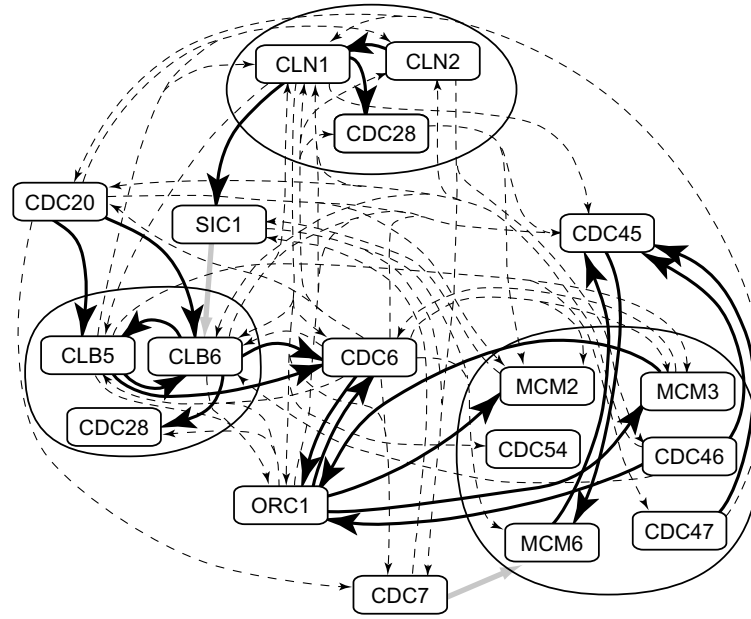
When we estimate networks from gene expression data alone, edge CLB6 → CDC6 in *S. cerevisiae* is estimated. This relationship also appears in KEGG pathways. Therefore, this is a correctly estimated edge. On the other hand, the corresponding relationship in *H. sapiens*, that is, CCNA1 → CDC6, is not estimated from gene expression data alone. These *H. sapiens* genes are also connected in the KEGG pathways. By using the proposed method, we observed that this relationship is correctly estimated. Therefore, this is a typical example of successfully estimated relationships by the evolutionary information.

(b) *S. cerevisiae* and *H. sapiens*: relationships between MCM and ORC complexes

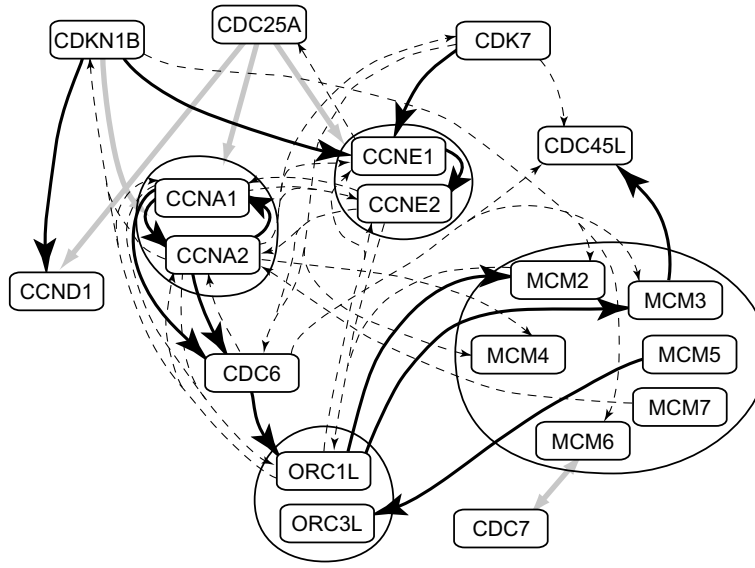
According to the KEGG pathways, MCM genes and ORC genes interact with each other. By using the evolutionary information, ORC1 and MCM3 of *S. cerevisiae* are newly connected. Therefore, the proposed method succeeded in estimating more known relationships. We observed that some relationships between ORC1 and MCM genes estimated from only gene expression data are not estimated by the proposed method. Moreover, not all combinations of orthologous genes between these genes are connected by using the evolutionary information. This shows that the expression data of *H. sapiens* do not contain information that these genes are related to each other. These inconsistent relationships may suggest the difference in the cellular systems between *S. cerevisiae* and *H. sapiens*.

(c) *S. cerevisiae* and *H. sapiens*: MCM complexes → CDC6

The KEGG pathways do not contain relationships between MCM genes and CDC6 in both organisms. However, it is known that these products interact directly with each other in the both organisms.[25,26] From microarray data alone, 2 edges, CDC6 → MCM2 and MCM3 → CDC6 are estimated only for *S. cerevisiae*. By using the evolutionary information, two edges for *S. cerevisiae* and one edge for *H. sapiens* are newly estimated between these genes. According to Schepers and Diffley,[25] and Méndez and Stillman,[26] CDC6 is required for the loading of MCM complexes. Therefore, although CDC6 and MCM are not connected in the KEGG pathways, these edges are correct relationships and the proposed method succeeded in estimating them.



S. cerevisiac



H. sapiens

Fig. 4. Gene networks estimated by the proposed method in Experiment I. The solid edges are consistent with the KEGG pathways. The dashed edges are not contained in the KEGG pathways. The gray edges are known relationships in the KEGG pathways, but are not estimated. The edges within the MCM complex are omitted.

Table 3. Newly estimated edges in Experiment I by the proposed method.

S. cerevisiae	*H. sapiens*
CLB5 → CDC6	CDC6 → ORC1L
ORC1 → CDC6	CCNA1 → CDC6
ORC1 → MCM3	CCNA2 → CCNA1
CDC6 → ORC1	CCNA1 → CCNA2
MCM3 → ORC1	MCM2 → ORC1L
CLN2 → CLN1	CDC6 → MCM3
CDC6 → MCM3	ORC1L → CCNA1

3.2. *Experiment II: KEGG all cell cycle regulated genes + their homologous genes*

Dataset: As a further analysis, we applied the proposed method to a larger dataset. Since this experiment was conducted for discovering unknown relationships rather than for evaluating the proposed method, we considered to include genes that are not in the KEGG pathways. At first, we collected 44 genes of *S. cerevisiae* and 38 genes of *H. sapiens*, from all 101 genes and 109 genes appearing in the KEGG pathways, respectively. A gene is chosen if it was labeled as "cell cycle-regulated" in Spellman *et al.*[18] and Whitfield *et al.*[19] Next, we added nine genes of *S. cerevisiae* and 25 genes of *H. sapiens*, that are cell cycle-regulated and orthologous to the genes in the KEGG pathways, but not appearing in the pathways. In summary, we collected 53 genes for *S. cerevisiae* and 62 genes for *H. sapiens* in this analysis (Table 4).

Result: Since many genes used in this experiment do not appear in the KEGG pathways, it is difficult to use KEGG information for evaluating the results of this analysis as in Experiment I. Instead of using the KEGG, we used annotations from the GO (gene ontology) hierarchy.[20] If two genes have related functions, the same annotations are expected to be assigned to these genes in the GO annotations. Note that, in gene networks, not all connected genes are assigned the same annotations, but we can expect that connected genes in the network tend to share the same GO annotations. For the comparison of the estimated network, we calculated the average number of annotations which are commonly assigned to connected genes in the network. For the networks estimated by using only microarray data, the average numbers of GO annotations assigned in common are 1.449 for *S. cerevisiae*, and 1.667 for *H. sapiens*. On the other hand, the proposed method gives 2.202 for *S. cerevisiae*, and 2.313 for *H. sapiens*. Hence, we can conclude that the edges estimated by the proposed method contain at least more known relationships than the networks estimated by microarray data alone. The hyperparameters of the method for this experiment were selected in the same way as the previous experiment since a part of the selected genes has known relationships in the KEGG pathways ($\zeta_{\mathrm{P}} = 2.37$, $\zeta_{\mathrm{N}} = 0.57$). The estimated networks and the lists of all assigned annotations are

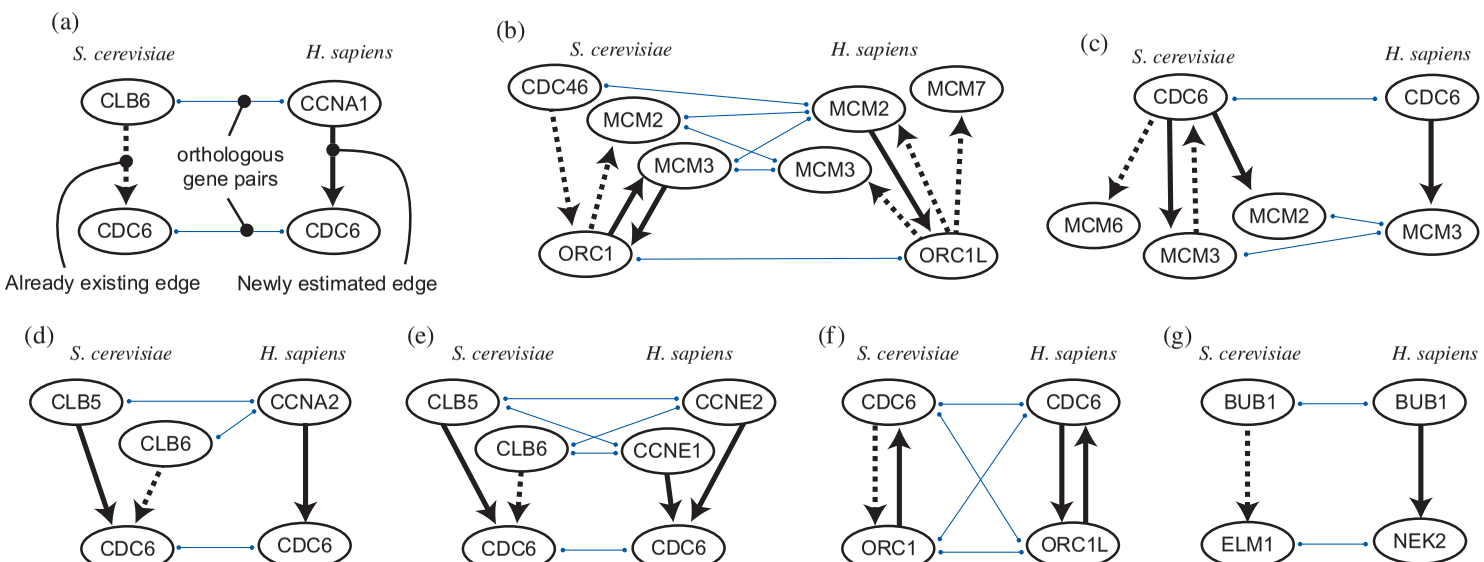


Fig. 5. Results of the real example. The dashed edges are estimated by both the previous method and the proposed method. The solid edges are estimated by the proposed method using the evolutionary information. The gene pairs that are connected by the lines are orthologous genes.

Table 4. List of genes in Experiment II. Gene names with underlines do not appear in the KEGG pathways, but are orthologous to the KEGG genes.

S. cerevisiae (53 genes)
APC1, BUB1, BUB2, CDC20, CDC45, CDC46, CDC47, CDC5, CDC54, CDC6, CLB1, CLB2, CLB4, CLB5, CLB6, CLN1, CLN2, CLN3, , DAM1, DBF2, DBF20, DUN1, ELM1, FAR1, GIN4, HSL1, HSL7, KIN3, MCD1, MCM2, MCM3, MCM6, MEC3, MOB1, ORC1, PCL1, PCL2, PDS1, PHO5, POL30, PRR1, RAD53, SCC3, SCH9, SIC1, SLT2, SMC1, SMC3, SWE1, SWI4, TEM1, VHS1, YCK1
H. sapiens (62 genes)
BUB1, BUB1B, BUB3, CCNA2, CCNB1, CCNB2, CCND1, CCNE1, CCNE2, CCNF, CDC16, CDC20, CDC25A, CDC25B, CDC25C, CDC27, CDC42, CDC45L, CDC6, CDC7, CDK7, CDKN1B, CDKN2C, CDKN2D, CENPE, CENPF, CIT, DKFZP434C245, DSP, E2F1, E2F5, FZR1, GADD45A, HDAC3, MAD2L1, MAP2K6, MAP3K2, MAPK13, MCM2, MCM4, MCM5, MCM6, MDM2, MPHOSPH1, NEK2, NKTR, ODF2, ORC1L, ORC3L, PCNA, PKMYT1, PLK, PTTG1, RAB23, RAB3A, RAD21, RAN, ROCK1, SGK, SMC4L1, STK17B, TTK

available in the supplementary information. Next, as in Experiment I, we evaluated newly estimated edges in the networks.

(d) *S. cerevisiae*: CLB5/6 → CDC6 and *H. sapiens*: CCNA2 → CDC6

It is known that CLB5/6 is required to start DNA replication in *S. cerevisiae*,[27] and this relationship also appears in the KEGG pathways. In *H. sapiens*, CCNA2 (Cyclin A2) is a corresponding gene to CLB5/6 in our orthologous gene data, and is also known to be required for the activity of *H. sapiens* CDC6,[28] and the KEGG pathways also contain this relationship. From gene expression data alone, CLB6 → CDC6 in *S. cerevisiae* is estimated while CCNA2 → CDC6 in *H. sapiens* is not. Our proposed method succeeded in estimating CCNA2 → CDC6 by utilizing the information of these orthologous genes. In addition to this relationship of *H. sapiens*, we observed that CLB5 → CDC6 in *S. cerevisiae* is newly estimated by the proposed method.

(e) *S. cerevisiae*: CLB5/6 → CDC6 and *H. sapiens*: CCNE1/2 → CDC6

As well as in the previous example, CCNE1 and 2 (Cyclin E) of *H. sapiens* are also orthologous genes of CLB5/6 of *S. cerevisiae*. Although the KEGG pathways do not contain relationships between CCNE1/2 and CDC6, it is known that CCNE1/2 contribute to the stabilization of CDC6.[29] Whereas this relationship is not estimated from gene expression data alone, the proposed method succeeded in detecting the relationship by utilizing information of corresponding orthologous genes.

(f) *S. cerevisiae*: CDC6 ↔ ORC1 and *H. sapiens*: CDC6 ↔ ORC1L

CDC6 is known to regulate genes involved in DNA replication, including ORC genes in eukaryotic cells. From microarray data alone, only CDC6 → ORC1 in *S. cerevisiae* is estimated. In our dataset, both ORC1L and CDC6 of *H. sapiens* are orthologous genes of *S. cerevisiae* ORC1 and CDC6. Although no relationship between ORC1L and CDC6 in *H. sapiens* is estimated from the expression data

alone, our proposed method estimated this relationship. Since the KEGG pathways contain these relationships, the estimated relationships can be considered to be consistent with biological knowledge. Because both ORC1 and CDC6 are orthologous to each other in our dataset, our proposed method estimated these relationships in both directions.

(g) *S. cerevisiae*: BUB1 → ELM1 and *H. sapiens*: BUB1 → NEK2

From gene expression data alone, a relationship from BUB1 to ELM1 is estimated in *S. cerevisiae*. NEK2 of *H. sapiens*, which is a corresponding gene to ELM1, is known to interact with MAD1.[30] In the KEGG pathways, BUB1 interacts with MAD1 (BUB1 → MAD1), and moreover, NEK2, MAD1 and BUB1 are known to play important roles for the spindle checkpoint function in M phase.[30,31] By our proposed method, BUB1 → NEK2 is newly estimated. Considering the fact that MAD1 is not selected for the estimation of the network, this BUB1 → NEK2 may be a correctly estimated relationship in *H. sapiens*. On the other hand, ELM1 of *S. cerevisiae* is known to function in a mitotic signaling network[32] in M phase, however, BUB1 and ELM1 are not known to interact with each other. According to the fact that the corresponding relationship in *H. sapiens* is correct and BUB1 also acts in M phase, BUB1 and ELM1 in *S. cerevisiae* may possibly be related to each other.

4. Conclusion

We have proposed a statistical framework for estimating simultaneously two gene networks from microarray gene expression data, utilizing the evolutionarily conserved relationships between two organisms. For evaluating the proposed method, we applied it to *S. cerevisiae* and *H. sapiens* cell cycle gene expression data, and confirmed that the information on the estimated network of an organism can be used to improve the accuracy of the network of the other organism. In Experiment I, whereas the previous method could only estimate a relatively small part of known relationships, especially in *H. sapiens*, our method drastically improved both the specificity and the sensitivity of the networks when compared to the known relationships in the KEGG pathways. In Experiment II, we applied our method to genes that do not appear in the KEGG pathways but are possibly related to the cell cycle, in addition to the genes in the KEGG pathways. In this experiment, we also succeeded in estimating much more known relationships than the previous method.

In both experiments, our proposed method estimated many unknown relationships that do not appear in the KEGG pathways but seem biologically plausible. Such observations may be worth confirming whether they represent actual relationships or not by biological experiments. For the further evaluation of the proposed method, computational experiments with other organisms and their gene expression data are preferable. However, there are only a few datasets where both detailed knowledge of the regulatory pathways and the same types of the expression data are

available for two distinct organisms. Therefore, we would like to apply our proposed method to such new datasets when they become available.

References

1. Van Someren EP, Genetic network modeling, *Pharmacogenomics* **3**:507–252, 2002.
2. Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S, Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, *J Bioinform Comput Biol* **2**:77–98, 2004.
3. Savoie CJ, Aburatani S, Watanabe S, Eguchi Y, Muta S, Imoto S, Miyano S, Kuhara S, Tashiro K, Use of gene networks from full genome microarray libraries to identify functionally relevant drug-affected genes and gene regulation cascades, *DNA Res* **10**:19–25, 2003.
4. Friedman N, Linial M, Nachmann I, Pe'er D, Using Bayesian network to analyze expression data, *J Comput Biol* **7**:601–620, 2000.
5. Ong IM, Glasner JD, Page D, Modelling regulatory pathways in *E. coli* from time series expression profiles, *Bioinformatics* **18**:S241–S248, 2002.
6. Smith VA, Jarvis ED, Hartemink AJ, Evaluating functional network inference using simulations of complex biological systems, *Bioinformatics* **18**:S216–S224, 2002.
7. Imoto S, Goto T, Miyano S, Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression, *Pac Symp Biocomput* **7**:175–186, 2002.
8. Imoto S, Kim S, Goto T, Aburatani S, Tashiro K, Kuhara S, Miyano S, Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *J Bioinform Comput Biol* **1**:231–252, 2003.
9. Kim S, Imoto S, Miyano S, Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems* **75**:57–65, 2004.
10. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA, Combining location and expression data for principled discovery of genetic regulatory network models, *Pac Symp Biocomput* **7**:437–449, 2002.
11. Nariai N, Kim S, Imoto S, Miyano S, Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks, *Pac Symp Biocomput* **9**:336–347, 2004.
12. Tamada Y, Kim S, Bannai H, Imoto S, Tashiro K, Kuhara S, Miyano S, Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection, *Bioinformatics* **19**:ii227–ii236, 2003.
13. Stuart JM, Segal E, Koller D, Kim SK, A gene-coexpression network for global discovery of conserved genetic modules, *Science* **302**:249–255, 2003.
14. Teichmann SA, Babu MM, Conservation of gene co-regulation in prokaryotes and eukaryotes, *TRENDS Biotechnol* **20**:407–410, 2002.
15. Teichmann SA, Babu MM, Gene regulatory network growth by duplication, *Nat Genet* **36**:492–496, 2004.
16. Snel B, van Noort V, Huynen MA, Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes, *Nucleic Acids Res* **32**:4725–4731, 2004.
17. Kanehisa M, Goto S, KEGG: Kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**:27–30, 2000.
18. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstien D, Futcher B, Comprehensive identification of cell cycle-regulated genes of the

Yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol Biol Cell* **9**:3273–3297, 1998.
19. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D, Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Mol Biol Cell* **13**:1977–2000, 2002.
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Gene ontology: tool for the unification of biology. The gene ontology consortium, *Nat Genet* **25**:25–29, 2000.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool, *J Mol Biol* **215**:403–410, 1990.
22. Babu MM, Teichmann SA, Evolution of transcription factors and the gene regulatory network in *Escherichia coli*, *Nucleic Acids Res* **31**:1234–1244, 2003.
23. Ott S, Miyano S, Finding optimal gene networks using biological constraints, *Genome Inform* **14**:124–133, 2003.
24. Ott S, Imoto S, Miyano S, Finding optimal models for small gene networks, *Pac Symp Biocomput* **9**:557–567, 2004.
25. Schepers A, Diffley JF, Mutational analysis of conserved sequence motifs in the budding yeast Cdc6 protein, *J Mol Biol* **308**:597–608, 2001.
26. Méndez J, Stillman B, Chromatin association of human origin recognition complex, CDC6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis, *Mol Cell Biol* **20**:8602–8612, 2000.
27. Toone WM, Aerne BL, Morgan BA, Johnston LH, Getting started: regulating the initiation of DNA replication in yeast, *Annu Rev Microbiol* **51**:125–149, 1997.
28. Yam CH, Fung TK, Poon RYC, Cyclin A in cell cycle control and cancer, *Cell Mol Life Sci* **59**:1317–1326, 2002.
29. Bermejo R, Vilaboa N, Calés C, Regulation of CDC6, geminin, and CDT1 in human cells that undergo polyploidization, *Mol Biol Cell* **13**:3989–4000, 2002.
30. Lou Y, Yao J, Zereshki A, Dou Z, Ahmed K, Wang H, Hu J, Wang Y, Yao X, NEK2A interacts with MAD1 and possibly functions as a novel integrator of the spindle checkpoint signaling, *J Biol Chem* **279**:20049–20057, 2004.
31. Brady DM, Hardwick KG, Complex formation between Mad1p, Bub1p and Bub3p is crucial for spindle checkpoint function. *Curr Biol* **10**:675–678, 2000.
32. Sreenivasan A, Kellogg D, The elm1 kinase functions in a mitotic signaling network in budding yeast, *Mol Cell Biol* **19**:7983–7994, 1999.

Yoshinori Tamada is currently a doctor course student of the Bioinformatics Center, Institute for Chemical Research, Kyoto University. He received his M.S. in Mathematical Science from Tokai University in 2003. His current research interests include developing computational methods for inferring gene networks from microarray gene expression data and other biological data.

Hideo Bannai is currently an Assistant Professor at the Department of Informatics, Kyushu University. He received his Ph.D. in Computer Science from the University of Tokyo in 2005. His current research interest concerns string algorithms in general, including the study of index structures for strings and algorithms for optimal pattern discovery.

Seiya Imoto is currently a Research Associate of the laboratory of DNA analysis, Human Genome Center, Institute of Medical Science, University of Tokyo. He received BS, MS, and Ph.D. in Mathematics from Kyushu University in 1996, 1998 and 2001, respectively. His current research interests cover statistical analysis of high dimensional data by Bayesian approach, DNA microarray gene expression data analysis, gene regulatory network analysis and computational drug target discovery.

Toshiaki Katayama is currently a Research Associate of Human Genome Center, Institute of Medical Science, University of Tokyo. He received M.S. in biology from Kyoto University in 1998. His current interests include genome sequence analysis, standardization and providing web service of the KEGG database such as KEGG API and KEGG DAS, and development of the BioRuby open source bioinformatics library.

Minoru Kanehisa is a Director and Professor of Bioinformatics Center, Institute for Chemical Research, Kyoto University and a Professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He received Ph.D. in physics from the University of Tokyo in 1976. His current interests include computational prediction of cellular functions by integrated analysis of genomic, chemical, and network information with KEGG database. He is a principal investigator of the Japanese Genome Informatics Project (1991–2000), and a president of the Japanese Society for Bioinformatics (1999–2003).

Satoru Miyano is a Professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He obtained his Ph.D. in Mathematics from Kyushu University in 1984. His research group is developing computational methods for inferring gene networks from microarray gene expression data and other biological data, e.g., protein-protein interactions, promoter sequences. The group also developed a software tool called Genomic Object Net for modeling and simulation of various biological systems. This software is now commercialized as Cell Illustrator. Currently, his research group is intensively working for developing the gene networks of human endothelial cell by knocking down hundreds of genes. With these technical achievements, his research direction is now heading toward the creation of Systems Pharmacology.