Estimating gene regulatory networks and protein–protein interactions of *Saccharomyces cerevisiae* from multiple genome-wide data

Naoki Nariai¹,*, Yoshinori Tamada², Seiya Imoto¹ and Satoru Miyano¹

¹Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan and ²Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan

**ABSTRACT**

**Motivation:** Biological processes in cells are properly performed by gene regulations, signal transductions and interactions between proteins. To understand such molecular networks, we propose a statistical method to estimate gene regulatory networks and protein–protein interaction networks simultaneously from DNA microarray data, protein–protein interaction data and other genome-wide data.

**Results:** We unify Bayesian networks and Markov networks for estimating gene regulatory networks and protein–protein interaction networks according to the reliability of each biological information source. Through the simultaneous construction of gene regulatory networks and protein–protein interaction networks of *Saccharomyces cerevisiae* cell cycle, we predict the role of several genes whose functions are currently unknown. By using our probabilistic model, we can detect false positives of high-throughput data, such as yeast two-hybrid data. In a genome-wide experiment, we find possible gene regulatory relationships and protein–protein interactions between large protein complexes that underlie complex regulatory mechanisms of biological processes.

**Contact:** nariai@ims.u-tokyo.ac.jp

1 INTRODUCTION

Many biological processes are carried out by interactions between proteins, RNA and DNA in living cells. Recently, high-throughput analyses enabled us to obtain genome-wide information, such as mRNA expression, protein–protein interactions, protein localizations and so on. A lot of attention has been focused on developing computational methods for extracting valuable information of molecular networks from such various types of genomic data.

Currently, statistical methods for estimating gene regulatory networks from genomic data are mainly based on DNA microarray data (Akutsu et al., 1999; Chen et al., 1999; Friedman et al., 2000; Hartemink et al., 2002; Imoto et al., 2002, 2003; Pe’er et al., 2001; Shmulevich et al., 2002). However, since information contained in microarrays is limited by their quality, noise and experimental errors, using only microarray data is not enough for estimating gene regulatory networks accurately. Therefore, the use of additional biological data is considered as a key to microarray data analyses. There are several works combining microarray data with biological knowledge, such as localization data (Hartemink et al., 2002), DNA sequences of promoter elements (Pilpel et al., 2001; Tamada et al., 2003) and transcriptional bindings of regulators (Bernard and Hartemink, 2005; De Hoon et al., 2004; Imoto et al., 2004; Segal et al., 2003a,c).

However, protein–protein interaction networks are mainly constructed based on protein–protein interaction data observed, such as yeast two-hybrid assays or tandem-affinity purification (TAP) experiments (Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001; Jeong et al., 2001; Uetz et al., 2000). However, protein–protein interaction data often contain some errors, and it is not easy to construct comprehensive protein–protein interaction networks from these interaction data alone. Therefore, using other genomic data, such as mRNA expression, functional databases and essentiality phenotypes, is considered to be effective for more accurate prediction of protein–protein interactions (Jansen et al., 2003).

In this paper, we propose a statistical method for estimating gene regulatory networks and protein–protein interaction networks simultaneously based on microarray data, protein–protein interactions, protein localizations, essentiality phenotypes and functional categories. Figure 1 shows a conceptual view of the proposed method. The model consists of three components: a gene regulatory network model (directed graph) based on Bayesian networks, a protein–protein interaction network model (undirected graph) represented by binary Markov networks and a structural connection between gene regulatory networks and protein–protein interaction networks. The last part realizes the connection between gene regulatory networks and protein–protein interaction networks, giving a penalty to coexistence of a directed edge and an undirected edge between genes. Since physically interacting proteins are often coexpressed (Ge et al., 2002), previous approaches often estimate the coexpressed relationship as a gene regulation instead of a protein–protein interaction. To overcome this drawback, we combine these three components as one statistical model under a Bayes statistics in order to distinguish gene regulations from protein–protein interactions clearly in the estimated network.

Previously, Segal et al. (2003b) proposed a clustering method for grouping genes that could be on the same pathway based on microarray data and protein–protein interaction data. In using protein–protein interaction information, they used binary information whether the protein–protein interaction is observed or not. However, the quality of each protein–protein interaction should be quantified according to its reliability. In our proposed method, we compute the reliability of protein–protein interactions and use this...
data which indicate protein–protein interactions between genes. Our
goal is to construct a gene regulatory network $G_r$ (directed graph)
and a protein–protein interaction network $G_p$ (undirected graph) that
maximize the joint posterior probability $P(G_r, G_p|X, Y)$. By remov-
ing the normalizing constant, we can decompose the joint posterior
probability as

$$P(G_r, G_p|X, Y) \propto P(G_r, G_p, X, Y)$$

$$= P(X|G_r)P(Y|G_p)P(G_r, G_p),$$

where $P(X|G_r, G_p) = P(X|G_r)$ and $P(Y|G_r, G_p) = P(Y|G_p)$
hold in our model. Here, $P(X|G_r)$ and $P(Y|G_p)$ show the likeli-
hoods of gene-expression data $X$ and protein–protein interaction data
$Y$ under given $G_r$ and $G_p$, respectively, and $P(G_r, G_p)$ shows
the joint prior probability of $G_r$ and $G_p$. That is, the proposed method
contains three components, $P(X|G_r)$, $P(Y|G_p)$ and $P(G_r, G_p)$,
and we elucidate how to construct them in the following sections.

### 2.1 Gene regulatory network model

Suppose that we have $n$ sets of microarray data $X = \{x_1, \ldots, x_n\}$
of $p$ genes. A Bayesian network gives a solution to compute $P(X|G_r)$
by using the structure of the directed acyclic graph, $G_r$, and assum-
ing the Markov relationship between nodes. By using a Bayesian
network, we have the decomposition of the joint probability based
on the graph, $G_r$:

$$P(y_{ij}|x, G_r) = \prod_{i=1}^{n} f_{ij}(x_{ij} | p_{ij}, \theta),$$

where $x_{ij}$ is the expression value of gene $i$, of $i$-th microarray, $p_{ij}$ is the
vector of expression values of the direct parents of gene $i$, of $i$-th micro-
array and $\theta$ is the parameter vector. The likelihood
of gene-expression data can be computed as

$$P(X|G_r) = \int \prod_{i=1}^{n} f(x_{ij} | \theta, G_r)\pi(\theta|G_r, \lambda)d\theta,$$

where $\pi(\theta|G_r, \lambda)$ is the prior distribution on the parameter $\theta$ and $\lambda$ is
the hyperparameter vector. In this paper, we use the non-parametric
regression model with $B$-splines (Imoto et al., 2002, 2003) for
constructing Bayesian networks.

### 2.2 Protein–protein interaction network model

As a measure of reliability for each protein–protein interaction,
Jansen et al. (2003) proposed to compute a likelihood ratio for each
protein pair. Let $y_{ij}(k)$ be an element of $Y$ that shows a genomic fea-
ture of protein pair, gene, and gene. For example, suppose that an
experiment corresponding to $k = 1$ is a yeast two-hybrid assay. Then
$y_{ij}(1) = 1$ (or 0) means that the protein pair of gene, and gene
interacted (or did not interact). The reliability of the protein–protein
interaction between gene, and gene is then given by the likelihood ratio
$L(i, j) = \frac{P(y_{ij}(1)|\text{pos})}{P(y_{ij}(1)|\text{neg})} \times \cdots \times \frac{P(y_{ij}(N)|\text{pos})}{P(y_{ij}(N)|\text{neg})}.$

Under a given undirected graph $G_p$, the likelihood of protein–protein
interaction information $Y$ can be computed by a binary Markov

**Fig. 1.** Conceptual view of the proposed method. Gene regulatory networks
and protein–protein interaction networks are learned simultaneously from
biological information to construct a protein–protein interaction network. In
addition, our aim is different from theirs in that we estimate gene
regulatory networks and protein–protein interaction networks of a
cell, whereas they tried to find co-functioning genes on the same
pathway. On the other hand, Nariai et al. (2004) proposed a method
for estimating regulatory relationships between genes represented as
directed edges based on microarray data and protein–protein interac-
tion data. However, whether the estimated causal relationships show
gene regulations or protein–protein interactions are difficult to under-
stand. In our model, we clearly discern gene regulatory relationships
(directed edges) and protein–protein interactions (undirected edges),
and the information of protein–protein interaction networks helps to
refine gene regulatory networks and vice versa.

For evaluating our method, we conduct two real applications: First,
we construct both gene regulatory networks and protein–protein
interaction networks of *Saccharomyces cerevisiae* cell cycle from
mutant expression data (Hughes et al., 2000), protein–protein inter-
action data (Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001;
Uetz et al., 2000), essentiality phenotypes (Giaever et al., 2002),
and the MIPS functional category database (Mewes et al., 2002). Our
results show that the estimated gene regulatory networks success-
fully find more complex regulatory relationships, and the estimated
protein–protein interaction networks are improved in terms of both
the accuracy and coverage of known protein–protein interactions,
compared with the previous method applied separately. We also
suggest possible biological roles of functionally unknown genes
based on the information of estimated gene regulatory networks and
protein–protein interaction networks. As a second experiment, we
perform a genome-wide analysis. We estimate gene regulations and
protein–protein interactions of 5335 genes and predict comprehens-
ive functional networks among large protein complexes. The details
of the real data analyses are described in Section 4.

## 2 PROBABILISTIC MODEL

Let $X$ be gene-expression data and $Y$ be protein–protein interac-
tion data that include physical interaction data and other biological

### 2.1 Gene regulatory network model

Suppose that we have $n$ sets of microarray data $X = \{x_1, \ldots, x_n\}$
of $p$ genes. A Bayesian network gives a solution to compute $P(X|G_r)$
by using the structure of the directed acyclic graph, $G_r$, and assum-
ing the Markov relationship between nodes. By using a Bayesian
network, we have the decomposition of the joint probability based
on the graph, $G_r$:

$$P(y_{ij}|x, G_r) = \prod_{i=1}^{n} f_{ij}(x_{ij} | p_{ij}, \theta),$$

where $x_{ij}$ is the expression value of gene $i$, of $i$-th microarray, $p_{ij}$ is the
vector of expression values of the direct parents of gene $i$, of $i$-th micro-
array and $\theta$ is the parameter vector. The likelihood
of gene-expression data can be computed as

$$P(X|G_r) = \int \prod_{i=1}^{n} f(x_{ij} | \theta, G_r)\pi(\theta|G_r, \lambda)d\theta,$$

where $\pi(\theta|G_r, \lambda)$ is the prior distribution on the parameter $\theta$ and $\lambda$ is
the hyperparameter vector. In this paper, we use the non-parametric
regression model with $B$-splines (Imoto et al., 2002, 2003) for
constructing Bayesian networks.

### 2.2 Protein–protein interaction network model

As a measure of reliability for each protein–protein interaction,
Jansen et al. (2003) proposed to compute a likelihood ratio for each
protein pair. Let $y_{ij}(k)$ be an element of $Y$ that shows a genomic fea-
ture of protein pair, gene, and gene. For example, suppose that an
experiment corresponding to $k = 1$ is a yeast two-hybrid assay. Then
$y_{ij}(1) = 1$ (or 0) means that the protein pair of gene, and gene
interacted (or did not interact). The reliability of the protein–protein
interaction between gene, and gene is then given by the likelihood ratio
$L(i, j) = \frac{P(y_{ij}(1)|\text{pos})}{P(y_{ij}(1)|\text{neg})} \times \cdots \times \frac{P(y_{ij}(N)|\text{pos})}{P(y_{ij}(N)|\text{neg})}.$

Under a given undirected graph $G_p$, the likelihood of protein–protein
interaction information $Y$ can be computed by a binary Markov
network model (Segal et al., 2003b)

\[
P(Y|G_p) = \frac{1}{Z_p} \exp \left( \sum_{e(i,j) \in G_p} \lambda e(i,j) \right)
\]

where \(e(i,j)\) is the undirected edge between gene, and gene \(\gamma, Z_\gamma\) is the normalizing constant and \(\alpha\) is the reliability degree parameter \((\alpha \geq 0)\). The \(\alpha\) controls the balance between microarray data and protein–protein interaction information.

2.3 Connection between gene regulatory networks

We decompose the joint probability \(P(G_r, G_p)\) as \(P(G_r|G_p)P(G_p)\), where \(P(G_r|G_p)\) is the prior probability of \(G_r\) conditional on \(G_p\), and \(P(G_p)\) is the prior probability of \(G_p\). From structural information of an undirected graph \(G_p\), we define a value for the directed edge from gene, to gene \(g_j = 1\) for \(e[i,j] \not\in G_p\) and 2 for \(e[i,j] \in G_p\). By using \(c_{ij}\), we define the prior probability of \(G_r\) under a given \(G_p\) as

\[
P(G_r|G_p) \propto \exp \left( - \sum_{e(i,j) \in G_r} \zeta_{e_{ij}} \right),
\]

where \(e(i,j)\) is the directed edge from gene, to gene \(\gamma_{\gamma} \zeta_i \zeta_j\) are parameters \((0 \leq \zeta_i \leq \zeta_j)\). That is, \(\zeta\) tunes the complexity of \(G_r\) and \(\zeta_j\) adds a penalty to the structure of \(G_r\) according to the lower prior probability of \(e(i,j)\) if \(e(i,j)\) is included in \(G_p\).

We construct the prior probability of \(G_p\) as

\[
P(G_p) \propto \exp \left( - \zeta_p \sum_{\substack{\{i,j\} \in G_p}} I(e[i,j] \in G_p) \right),
\]

where \(\zeta_p\) is a complexity parameter \((\zeta_p \geq 0)\) that controls the complexity of \(G_p\), and \(I(e[i,j] \in G_p) = 1\) for \(e[i,j] \in G_p\) and 0 for \(e[i,j] \not\in G_p\). Hence, from Equations (5) and (6), the joint prior probability of gene regulatory networks and protein–protein interaction networks is defined by

\[
P(G_r, G_p) = \frac{1}{Z_{prior}} \exp \left( - \sum_{e(i,j) \in G_r} \zeta_{e_{ij}} - \zeta_p \sum_{\substack{\{i,j\} \in G_p}} I(e[i,j] \in G_p) \right),
\]

where \(Z_{prior}\) is the normalizing constant.

3 CRITERION AND ALGORITHM FOR ESTIMATING NETWORKS

We choose the graph structures of gene regulatory networks and protein–protein interaction networks by maximizing the joint posterior probability (1). For computing the integration in Equation (2), we used the Laplace approximation for integrals (Imoto et al., 2002; Konishi et al., 2004). Hence, we have a criterion, named GPNC (Gene regulatory networks and protein–protein interaction networks criterion) for evaluating gene regulatory networks and protein–protein interaction networks from Equations (2), (4) and (7) as:

\[
GPNC(G_r, G_p) = -2 \log \left( \frac{P(X|G_r)P(Y|G_p)P(G_r, G_p)}{P(G_r|G_p)P(G_p)} \right)
\]

\[
= -2 \log \int f(X|\theta, G_r)\pi(\theta|G_r, \lambda) d\theta + 2 \sum_{e(i,j) \in G_r} \zeta_{e_{ij}} - 2 \sum_{\{i,j\} \in G_p} \{\alpha \log L(i,j) - \zeta_p\} + Z,
\]

where \(f(X|\theta, G_r) = \prod_x f(x_i|\theta, G_r)\), and \(Z\) is the constant. The optimal \(G_r\) and \(G_p\) are chosen as the minimizers of Equation (8).

Based on the joint probabilistic model and the criterion described above, we use a greedy hill-climbing algorithm for estimating the gene regulatory network \(G_r\) and the protein–protein interaction network \(G_p\) under given parameters \(\alpha, \zeta_1, \zeta_2\) and \(\zeta_p\) as follows:

Step 1. Estimate \(G_p\) based on

\[
P(G_p|Y) \propto P(Y|G_p)P(G_p).
\]

Step 2.

Step 2-1. For gene \(y\), perform one of four procedures: ‘add a parent’, ‘remove a parent’, ‘reverse the parent–child relationship’ or ‘none’, which gives the lowest score. Update \(G_r\) and \(G_p\).

Step 2-2. If the score becomes unchanged, the learning is finished. Otherwise, go to Step 2 and continue the algorithm.

It is natural to consider that estimated causal relationships within protein complexes are protein–protein interactions. Therefore, after learning is finished, directed edges in protein complexes are changed to undirected edges. For example, if the directed edge from gene, to gene \(y\), exists in \(G_r\), but these two genes are connected in \(G_p\), we change the directed edge from gene, to gene \(y\), to undirected edge.

4 COMPUTATIONAL EXPERIMENT

4.1 Data preparation and parameter selection

For constructing protein–protein interaction networks, we collected protein–protein interaction data from four different experiments (Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001; Uetz et al., 2000), essentiality phenotypes (Giaever et al., 2002) and the MIPS functional category database (Mewes et al., 2002). We extract 9928 binary protein–protein interactions from the MIPS complex catalogue (Mewes et al., 2002) for constructing the positive interaction pairs, and extract 14 224 045 different localizing pairs from the MIPS localization data (Mewes et al., 2002) for constructing negatives. Within 9928 positive protein pairs, 428 protein pairs also belong to the negatives. However, as the fraction of 4% is small and some proteins localize differently in different biological processes, we consider that the positives and negatives we constructed serve as a good practical approximation.

Table 1 shows the likelihood ratios of all 16 combinations of the binary protein–protein interactions from four different experiments described above. Next, Table 2 shows the likelihood ratios of essential phenotypes. If two proteins are included in a biologically
Table 1. The likelihood ratio of protein–protein interactions

<table>
<thead>
<tr>
<th>G</th>
<th>H</th>
<th>I</th>
<th>U</th>
<th>Number of pairs</th>
<th>pos</th>
<th>neg</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>Inf.</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>17</td>
<td>4</td>
<td>0</td>
<td>Inf.</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>28</td>
<td>19</td>
<td>3</td>
<td>9073.9</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>5</td>
<td>1</td>
<td>7163.6</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>49</td>
<td>27</td>
<td>7</td>
<td>5526.2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>33</td>
<td>12</td>
<td>4</td>
<td>4298.2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>6</td>
<td>4</td>
<td>2149.1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1573</td>
<td>364</td>
<td>355</td>
<td>1469.0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>43</td>
<td>6</td>
<td>12</td>
<td>716.4</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>29</td>
<td>4</td>
<td>9</td>
<td>636.8</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>111</td>
<td>14</td>
<td>48</td>
<td>417.9</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16 130</td>
<td>1323</td>
<td>5525</td>
<td>343.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>670</td>
<td>7</td>
<td>326</td>
<td>30.8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>29 269</td>
<td>147</td>
<td>12669</td>
<td>16.6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4115</td>
<td>23</td>
<td>2556</td>
<td>12.9</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20 182 230</td>
<td>7960</td>
<td>14 202 526</td>
<td>0.8</td>
</tr>
</tbody>
</table>

G, H, I and U in this table show protein–protein interactions observed by Gavin et al., Ho et al., Ito et al. and Uetz et al., respectively. For example, if gene i and gene j have a protein–protein interaction observed by Gavin et al. and not by others, \( L(i, j) = \frac{p(1323)}{p(9285)} = 343.1 \).

Table 2. The likelihood ratio of essential phenotypes

<table>
<thead>
<tr>
<th>Phenotypes</th>
<th>Number of pairs</th>
<th>pos</th>
<th>neg</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>606 651</td>
<td>1390</td>
<td>318 925</td>
<td>6.2</td>
</tr>
<tr>
<td>EN</td>
<td>5 796 520</td>
<td>2504</td>
<td>3 841 414</td>
<td>0.9</td>
</tr>
<tr>
<td>NN</td>
<td>13 831 170</td>
<td>6034</td>
<td>10 063 706</td>
<td>0.9</td>
</tr>
</tbody>
</table>

EE: Both genes are essential, EN: Only one gene is essential, NN: Both genes are not essential.

Table 3. The likelihood ratio of the functional category

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of pairs</th>
<th>pos</th>
<th>neg</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same</td>
<td>381 587</td>
<td>9340</td>
<td>167 483</td>
<td>79.9</td>
</tr>
<tr>
<td>Otherwise</td>
<td>19 852 754</td>
<td>588</td>
<td>14 056 562</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Finding optimal values of four parameters \( \alpha, \zeta_1, \zeta_2 \) and \( \zeta_p \) in Equation (8) is intractable even for the moderate number of genes, because we need to compute the normalizing constants in Equations (4) and (7). To solve this problem, we simplify our model as follows: gene regulatory networks and protein–protein interaction networks are mutually exclusive and we assume no prior information on \( G_r \), i.e. we formally set \( \zeta_1 = 0 \) and \( \zeta_2 = \infty \). Since physical protein–protein interactions should be considered as protein–protein interaction networks instead of gene regulatory networks, we consider this assumption to be appropriate in practice. From Equation (8), \( \alpha \) and \( \zeta_p \) are included in \( \sum_{e \in G_p} \alpha \log \frac{L(i, j)}{L'} - \zeta_p \). By transforming \( \zeta_p \) to \( \alpha \log L' \), this term results in \( \alpha \sum_{e \in G_p} \log \frac{L(i, j)}{L'} \). Therefore, we consider \( \alpha \) and \( L' \) as parameters and set the candidate values as: \( L' = \{1, 300, 600, 900, 1200\} \) and \( \alpha = \{0.01, 0.1, 1, 10, 100\} \). To avoid local minima of the greedy algorithm, we repeated our algorithm 10 times for each parameter set, and then selected one network that gave the smallest score.
4.2 Cell-cycle network

We chose 297 genes of *S. cerevisiae* that are listed as cell-cycle related by Spellman *et al.* (1998). We used 56 cell-cycle related disruptant microarrays from 300 diverse mutations and chemical treatments (Hughes *et al*., 2000) by considering annotations of the MIPS database (Mewes *et al*., 2002). The number of selected genes was reduced to 290, considering the missing values of the microarrays.

After estimating gene regulatory networks and protein–protein interaction networks for specified $L'$ and $\alpha$, we counted the number of known regulatory relationships estimated in $G_p$. We collected 204 regulatory relationships from the location binding experiment by Lee *et al.* (2002) ($p$-value $\leq 0.05$) and considered them as known regulatory relationships. We suppose that gene, and gene, have a regulatory relationship if two genes are connected by a directed path in $G_p$, whose distance is within 2. Figure 2 shows the number of known regulatory relationships estimated by our method. We chose the most appropriate $\alpha$ for each $L'$ so that the maximum number of known regulatory relationships is estimated in $G_p$. Compared with the gene regulatory network estimated from microarray data alone, we successfully found more known regulatory relationships by adding the information of $G_p$.

For evaluating the estimated protein–protein interaction networks, we computed the accuracy ($T_p/P$) and coverage ($T_p/T$) of the protein–protein interaction network, where $T_p$ is the number of known protein–protein interactions estimated in $G_p$, $P$ is the number of all undirected edges in $G_p$, and $T$ is the number of known protein–protein interactions among 290 genes. Figures 3 and 4 show the accuracy and coverage of the estimated protein–protein interaction networks, respectively. Note that $\alpha$ for each $L'$ is the same as in Figure 2. We observe that both the accuracy and coverage of the estimated protein–protein interaction network are improved when $L' = 600$ and $\alpha = 1$, compared with the method without using the information of $G_p$. From this result, $L' = 600$ can be considered as a kind of threshold for the likelihood ratio defined in Equation (3). According to Jansen *et al.* (2003), the prior odds defined by $P(\pos)/P(\neg)$ is about 1/600. Therefore, our choice of $L'$ seems to be reasonable. Note that although there were 20 protein–protein interactions observed by yeast two-hybrid assays (Uetz *et al*., 2000; Ito *et al*., 2001) among 290 cell-cycle related genes, only 9 interactions were estimated as the protein–protein interactions in $G_p$. Among the nine interactions, four interactions were also observed by other experiments (Gavin *et al*., 2002; Ho *et al*., 2002). On the contrary, among the 11 interactions that were not estimated in $G_p$, only 1 interaction was also observed by another experiment (Ho *et al*., 2002). This result suggested that our method successfully reduced false positives of yeast two-hybrid assays. However, among all 105 estimated protein–protein interactions, 95 interactions have at least one physical interaction in Table 1. This result indicates that essential phenotypes and functional category information are only weak indicators of protein–protein interactions, compared with physical interactions. However, these information act as supporting data, which strengthen or weaken the reliability of protein–protein interactions.

Figure 5 shows three connected components in the estimated $G_p$. Within these components, we predict several protein–protein interactions that are not included in physical protein–protein interaction datasets (bold edges). POL1 and POL2 are catalytic subunits of DNA polymerase $\alpha$ and DNA polymerase $\epsilon$, respectively, and these two DNA polymerases work together. HHT1, HTA2, HTB1, HHF1 and HHF2 are histone genes, and MCM3 and MCM6 are subunits of a MCM complex. These facts support our findings of protein–protein interactions.

Figure 6 shows a part of the estimated gene regulatory networks and protein–protein interaction networks of *S. cerevisiae* cell cycle. We omit undirected edges within MCM and SMC complexes. We place genes on appropriate subcellular regions according to the MIPS localization data. Note that some genes in Figure 6 change their localizations at different biological phases. For example, Clb2p (M cyclin) is localized in the nucleus, cytoskeleton or mother-bud neck.
at appropriate cell-cycle stages. Similarly, protein–protein interactions represented by undirected edges are also condition specific. For example, Swe1p inhibits the activity of Clb2-Cdc28p by phosphorylation at G1/S phase. However, at G2/M phase, Hsl1p and Hsl7p promote the Swe1p degradation (McMillan et al., 1999), and hence the interaction between Swe1p and Clb2-Cdc28p is disappeared. Interestingly, the estimated network in Figure 6 reflects these regulatory relationships quite well, despite our network model not taking environmental conditions into account.

TOF2 and VIK1 (denoted by red circles in Figure 6) are still functionally unknown. TOF2 has a high sequence similarity to NET1 (E-value = 6.3 × 10^{-27}), and it was previously reported that Cdc5p influenced phosphorylation of Net1p (Shou and Deshaies, 1999). Interestingly, in our estimated network, there is a directed edge from CDC5 to TOF2. Hence, the estimated network could suggest that a possible biological role of TOF2 is similar to NET1 (regulator of nucleolar silencing and telophase exit). However, VIK1 has a high sequence similarity to CIK1 (E-value = 1.3 × 10^{-23}), but these genes seem to have different functions. Kar3-Cik1p attends the chromosome segregation, whereas Kar3-Vik1p attends the microtubule depolymerizing activity that opposes the spindle pole body separating force generated by Cin8p (Manning et al., 1999). It seems that our estimated network captures the different roles of Cik1p and Vik1p correctly, suggesting that the network contains other biologically meaningful relationships.

### 4.3 Genome-wide analysis

We apply our method to estimate a genome-wide network of *S. cerevisiae*. We used all 300 microarrays and selected 5335 genes for the analysis by considering missing values of the microarrays. For the setting of the parameters, we use $L' = 600$ and $a = 1$ as in Section 4.2. Since the number of genes is large and the estimated network becomes quite complicated, we evaluate the estimated network in the sense of intercomplex networks, i.e. we analyze gene regulatory networks and protein–protein interactions between protein complexes.

Figure 7 shows the intercomplex network of 10 protein complexes extracted from the estimated network by the following steps: First, we consider protein complexes in the MIPS complex catalog as positives and selected the 10 largest protein complexes overlapping with estimated $G_p$. Gray lines indicate the estimated protein–protein interactions in the positive protein complexes. Red lines and green lines indicate gene regulations and protein–protein interactions between complexes, respectively. Labels on the lines show the total number of each estimated edge.

From Figure 7, we observe that there are particularly many protein–protein interactions between 19/22S regulators and 20S proteasomes, between prereplication complex and replication complex, and among RNA polymerases I, II and III. As 19S regulators and 20S proteasomes constitute 26S proteasomes, 69 protein–protein interactions between these complexes seem biologically plausible. Similarly, many protein–protein interactions between prereplication complexes and replication complexes, and among RNA polymerases I, II and III seem natural because these complexes have similar functions. However, there are comparatively weak protein–protein interactions, such as between 19/22S regulators and prereplication complexes, and among RNA polymerase II, TAFIIs and SRB complexes. Since the role of 19S regulators is to unfold the protein substrates and inject them into the 20S proteasome for degradation, protein–protein interactions between 19/22S regulators and prereplication complexes would happen when prereplication complexes are degraded at appropriate cell-cycle phases. From these results, we could conclude that we successfully estimated plausible protein–protein interactions among these protein complexes.

Interestingly, several gene regulations were estimated between replication complexes and 20S proteasomes, and between rRNA splicing complexes and RNA polymerases I, II and III, while protein–protein interactions are not present between them except one protein–protein interaction between rRNA splicing complexes and RNA polymerase III (the green line is omitted in Figure 7). Since replication complexes are degraded at appropriate cell-cycle phases by the proteasome, and the cellular processes of RNA splicing are strongly linked to RNA polymerization, estimated gene regulatory relationships between these complexes would be meaningful in the biological sense. Note that in our gene regulatory networks model, causal relationships between genes estimated from microarray data are not necessarily transcriptional gene regulations. For example, it might be a case that some of the estimated gene regulations are not transcriptional.
regulatory relationships between 19/22S regulators and 20S proteasomes might be protein–protein interactions, as there are many protein–protein interactions estimated between them. This result indicates that more protein–protein interaction data are needed for distinguishing between physical interactions and other regulatory interactions correctly.

We can conclude that by estimating both gene regulatory networks and protein–protein interaction networks, we successfully obtain comprehensive functional networks among the 10 protein complexes.

5 DISCUSSION
In this paper we proposed a probabilistic model for estimating both gene regulatory networks and protein–protein interaction networks based on microarray data, protein–protein interactions and other genome-wide data. An example of *S.cerevisiae* cell-cycle related network showed that we successfully estimated gene regulatory networks and protein–protein interaction networks more accurately than the previous methods applied separately, and the estimated network suggested biological roles of functionally unknown genes. In a genome-wide analysis, we predicted comprehensive functional networks of 10 protein complexes by estimating both gene regulatory networks and protein–protein interaction networks. We consider the following topics as our future works: first, our current algorithm for learning gene regulatory networks and protein–protein interaction networks remains to be improved, as it is difficult to find optimal networks simply by greedy hill-climbing. Second, because it is important to know which gene regulations or protein–protein interactions are activated and under which conditions, we need to incorporate environmental conditions into our network model. Ideker *et al.* (2002) proposed a method for identifying active subnetworks in a molecular interaction network under a particular condition. Finally, in our current gene regulatory networks model, directed edges might include signal transductions, phosphorylations, ubiquitinations and so on, other than transcriptional gene regulations. For more accurate estimation of gene regulatory networks, we would better include the prior information, such as DNA sequences of promoter elements and DNA bindings of regulators. These are currently the limitations of the proposed approach and would be our future works.

We expect that an increasing number of microarray data and protein–protein interaction data enable us to analyze a broad range of biological processes, and elucidating both their gene regulatory networks and protein–protein interaction networks is a key to understand the complex nature of cellular functions.

Conflict of Interest: none declared.

REFERENCES


