BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii206–ii212
doi:10.1093/bioinformatics/bti1133

Systems Biology

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

Naoki Nariai[1,*], Yoshinori Tamada[2], Seiya Imoto[1] and Satoru Miyano[1]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan and [2]Bioinformatics 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

*To whom correspondence should be addressed.

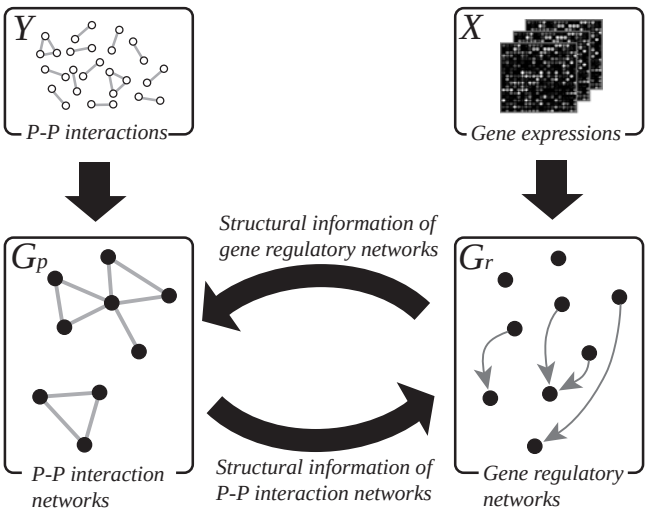

Fig. 1. Conceptual view of the proposed method. Gene regulatory networks and protein–protein interaction networks are learned simultaneously from biological data.

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 interaction data. However, whether the estimated causal relationships show gene regulations or protein–protein interactions are difficult to understand. 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 interaction 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 successfully find more known 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 comprehensive functional networks among large protein complexes. The details of the real data analyses are described in Section 4.

2 PROBABILISTIC MODEL

Let $\boldsymbol{X}$ be gene-expression data and $\boldsymbol{Y}$ be protein–protein interaction data that include physical interaction data and other biological 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 $P(G_r, G_p|\boldsymbol{X}, \boldsymbol{Y})$. By removing the normalizing constant, we can decompose the joint posterior probability as

$$P(G_r, G_p|\boldsymbol{X}, \boldsymbol{Y}) \propto P(G_r, G_p, \boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{X}|G_r)P(\boldsymbol{Y}|G_p)P(G_r, G_p), \quad (1)$$

where $P(\boldsymbol{X}|G_r, G_p) = P(\boldsymbol{X}|G_r)$ and $P(\boldsymbol{Y}|G_r, G_p) = P(\boldsymbol{Y}|G_p)$ hold in our model. Here, $P(\boldsymbol{X}|G_r)$ and $P(\boldsymbol{Y}|G_p)$ show the likelihoods of gene-expression data $\boldsymbol{X}$ and protein–protein interaction data $\boldsymbol{Y}$ 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, $P(\boldsymbol{X}|G_r)$, $P(\boldsymbol{Y}|G_p)$ 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 $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of p genes. A Bayesian network gives a solution to compute $P(\boldsymbol{X}|G_r)$ by using the structure of the directed acyclic graph, G_r, and assuming the Markov relationship between nodes. By using a Bayesian network, we have the decomposition of the joint probability based on the graph, G_r: $f(\boldsymbol{x}_i|\boldsymbol{\theta}, G_r) = \prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij}, \boldsymbol{\theta}_j)$, where x_{ij} is the expression value of gene$_j$ of i-th microarray, $\boldsymbol{p}_{ij}$ is the vector of expression values of the direct parents of gene$_j$ of i-th microarray and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_p')'$ is the parameter vector. The likelihood of gene-expression data can be computed as

$$P(\boldsymbol{X}|G_r) = \int \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta}, G_r)\pi(\boldsymbol{\theta}|G_r, \boldsymbol{\lambda})d\boldsymbol{\theta}, \quad (2)$$

where $\pi(\boldsymbol{\theta}|G_r, \boldsymbol{\lambda})$ is the prior distribution on the parameter $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ 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 $\boldsymbol{Y}$ that shows a genomic feature of protein pair, gene$_i$ and gene$_j$. 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$_i$ and gene$_j$ interacted (or did not interact). The reliability of the protein–protein interaction between gene$_i$ and gene$_j$ is then given by the likelihood ratio $L(i, j) = P(y_{ij}(1), \ldots, y_{ij}(N)|\text{pos})/P(y_{ij}(1), \ldots, y_{ij}(N)|\text{neg})$, where 'pos' and 'neg' are respectively the positive and negative sets of protein pairs constructed in advance, and N is the number of genomic features we considered. We explain how to construct the positive and negative sets in Section 4. If each genomic feature is conditionally independent (for example, protein–protein interaction data and functional category database can be considered as independent information sources), $L(i, j)$ can be rewritten as

$$L(i, j) = \frac{P(y_{ij}(1)|\text{pos})}{P(y_{ij}(1)|\text{neg})} \times \cdots \times \frac{P(y_{ij}(N)|\text{pos})}{P(y_{ij}(N)|\text{neg})}. \quad (3)$$

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

network model (Segal *et al*., 2003b)

$$P(\boldsymbol{Y}|G_p) = \frac{1}{Z_y} \prod_{e\{i,j\}\in G_p} L(i,j)^{\alpha}, \quad (4)$$

where $e\{i,j\}$ is the undirected edge between gene$_i$ and gene$_j$, Z_y is the normalizing constant and α is the reliability degree parameter ($\alpha \geq 0$). The α controls the balance between microarray data and protein–protein interaction information.

2.3 Connection between gene regulatory networks and protein–protein interaction networks

We decompose the joint probability $P(G_r, G_p)$ as $P(G_r, G_p) = P(G_r|G_p)P(G_p)$, 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$_i$ to gene$_j$ by $c_{ij} = 1$ for $e\{i,j\} \notin G_p$ 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

$$P(G_r|G_p) \propto \exp\left(- \sum_{e(i,j)\in G_r} \zeta_{c_{ij}} \right), \quad (5)$$

where $e(i,j)$ is the directed edge from gene$_i$ to gene$_j$, ζ_1 and ζ_2 are parameters ($0 \leq \zeta_1 \leq \zeta_2$). That is, ζ_1 tunes the complexity of G_r and ζ_2 adds a penalty to the structure of G_r according to the information of G_p. By using the prior probability (5), we put the lower prior probability to $e(i,j)$ if $e\{i,j\}$ is included in G_p.

We construct the prior probability of G_p as

$$P(G_p) \propto \exp\left[- \zeta_p \sum_{i,j} I(e\{i,j\} \in G_p) \right], \quad (6)$$

where ζ_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 $e\{i,j\} \notin G_p$. Hence, from Equations (5) and (6), the joint prior probability of gene regulatory networks and protein–protein interaction networks is defined by

$$P(G_r, G_p) = \frac{1}{Z_{\text{prior}}} \exp\left\{ - \sum_{e(i,j)\in G_r} \zeta_{c_{ij}} - \zeta_p \sum_{i,j} I(e\{i,j\} \in G_p) \right\}, \quad (7)$$

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:

$$\begin{aligned} \text{GPNC}(G_r, G_p) &= -2\log P(\boldsymbol{X}|G_r)P(\boldsymbol{Y}|G_p)P(G_r, G_p) \\ &= -2\log \int f(\boldsymbol{X}|\boldsymbol{\theta}, G_r)\pi(\boldsymbol{\theta}|G_r, \boldsymbol{\lambda})\, d\boldsymbol{\theta} \\ &\quad + 2 \sum_{e(i,j)\in G_r} \zeta_{c_{ij}} \\ &\quad - 2 \sum_{e\{i,j\}\in G_p} \{\alpha \log L(i,j) - \zeta_p\} + Z, \end{aligned} \quad (8)$$

where $f(\boldsymbol{X}|\boldsymbol{\theta}, G_r) = \prod_i f(\boldsymbol{x}_i|\boldsymbol{\theta}, G_r)$, 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 α, ζ_1, ζ_2 and ζ_p as follows:

Step 1. Estimate $\tilde{G}_p$ based on

$$P(G_p|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|G_p)P(G_p).$$

Step 2.

- *Step 2-1*. For gene$_i$, perform one of four procedures: 'add a parent', 'remove a parent', 'reverse the parent–child relationship' or 'none', which gives the lowest score. Update $\tilde{G}_r$ and $\tilde{G}_p$.
- *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$_i$ to gene$_j$ exists in $\tilde{G}_r$ but these two genes are connected in $\tilde{G}_p$, we change the directed edge from gene$_i$ to gene$_j$ to the 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

G	H	I	U	Number of pairs	pos	neg	L
1	1	1	1	9	7	0	Inf.
0	1	1	1	17	4	0	Inf.
1	1	1	0	28	19	3	9073.9
1	1	0	1	15	5	1	7163.6
1	0	1	0	49	27	7	5526.2
1	0	0	1	33	12	4	4298.2
1	0	1	1	20	6	4	2149.1
1	1	0	0	1573	364	355	1469.0
0	1	1	0	43	6	12	716.4
0	1	0	1	29	4	9	636.8
0	0	1	1	111	14	48	417.9
1	0	0	0	16 130	1323	5525	343.1
0	0	0	1	670	7	326	30.8
0	1	0	0	29 269	147	12 669	16.6
0	0	1	0	4115	23	2556	12.9
0	0	0	0	20 182 230	7960	14 202 526	0.8

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, $L(i,j) = \frac{P(1323|\text{pos})}{P(5525|\text{neg})} = \frac{1323/9928}{5525/14224045} = 343.1$.

Table 2. The likelihood ratio of essential phenotypes

Phenotypes	Number of pairs	pos	neg	L
EE	606 651	1390	318 925	6.2
EN	5 796 520	2504	3 841 414	0.9
NN	13 831 170	6034	10 063 706	0.9

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

Category	Number of pairs	pos	neg	L
Same	381 587	9340	167 483	79.9
Otherwise	19 852 754	588	14 056 562	0.1

essential protein complex, deletion mutants of each protein are likely to produce a lethal phenotype. Finally, Table 3 shows the likelihood ratios of the functional category. If two proteins have the same biological function, they have a tendency to form a protein complex. We use the MIPS functional catalogue to find pairs performing the same function. Note that if two proteins appear together in at least one functional category, we regard this pair as having the same function.

Finding optimal values of four parameters α, ζ_1, ζ_2 and ζ_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

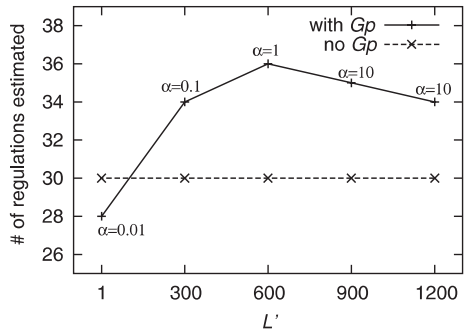

Fig. 2. The number of known regulatory relationships estimated by our method. Labels on the plots show the values of α where the maximum number of regulations was estimated.

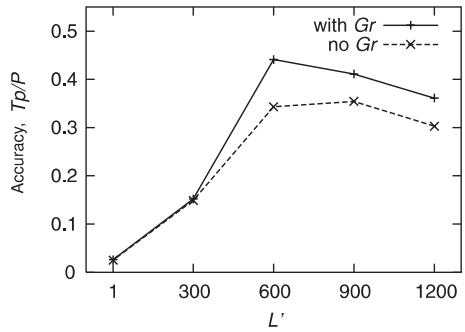

Fig. 3. The accuracy (T_p/P) of the estimated protein–protein interaction network.

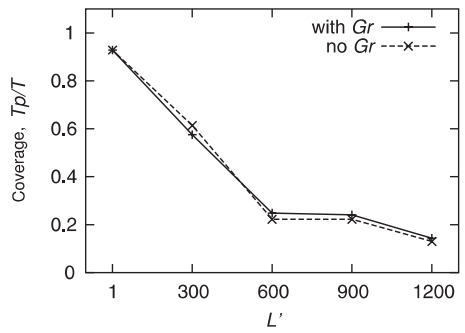

Fig. 4. The coverage (T_p/T) of the estimated protein–protein interaction network.

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), α and ζ_p are included in $\sum_{e\{i,j\}\in G_p}\{\alpha \log L(i,j) - \zeta_p\}$. By transforming ζ_p to $\alpha \log L'$, this term results in $\alpha \sum_{e\{i,j\}\in G_p} \log\{L(i,j)/L'\}$. Therefore, we consider α 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.

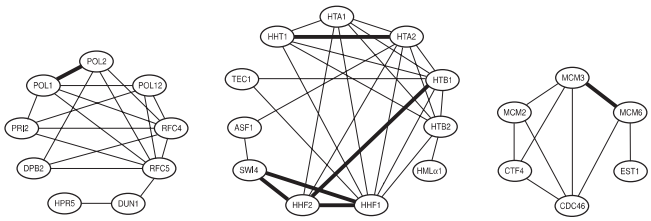

Fig. 5. Three connected components of the estimated protein–protein interaction networks. Bold edges in the graph indicate newly estimated protein–protein interactions that are not included in physical protein–protein interaction dataset.

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 α, we counted the number of known regulatory relationships estimated in G_r. We collected 204 regulatory relationships from the location binding experiment by Lee *et al*. (2002) (p-value ≤ 0.05) and considered them as known regulatory relationships. We suppose that gene$_i$ and gene$_j$ have a regulatory relationship if two genes are connected by a directed path in G_r whose distance is within 2. Figure 2 shows the number of known regulatory relationships estimated by our method. We chose the most appropriate α for each L' so that the maximum number of known regulatory relationships is estimated in G_r. 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 α 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_r. 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 $P(\text{pos})/P(\text{neg})$ 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.

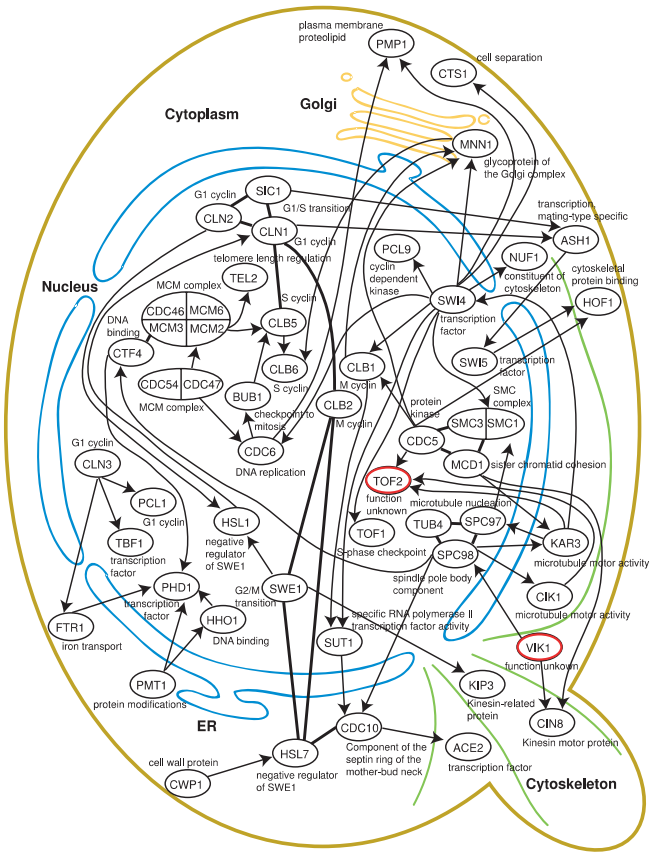

Fig. 6. *Saccharomyces cerevisiae* cell-cycle network estimated by the proposed method. Genes are located according to their localization.

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 α 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_r. 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 $P(\text{pos})/P(\text{neg})$ 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 α and DNA polymerase ε, 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

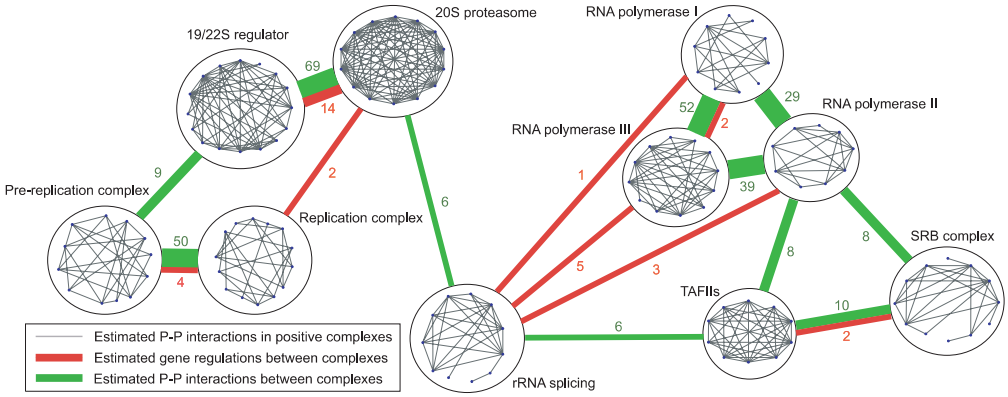

Fig. 7. Intercomplex network of 10 protein complexes estimated by the proposed method. Gray lines indicate the overlap of estimated protein–protein interactions and positive protein complexes. Red lines and green lines indicate estimated gene regulations and protein–protein interactions between complexes, respectively. Labels on the lines show the total number of each estimated edge.

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 G_1/S phase. However, at G_2/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 (BLAST E-value $= 6.3 \times 10^{-27}$), and it was previously reported that Cdc5p influenced phosphorylation of Net1p (Shou and Deshaies, 2002). 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 $= 1.3 \times 10^{-23}$), 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 $\alpha = 1$ 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. Note that in Figure 7, we consider the green lines whose label has the number >5 to be significant and other green lines are omitted.

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

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

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

Bernard,A. and Hartemink,A. (2005) Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. *Pac. Symp. Biocomput.*, **10**, 459–470.

Chen,T. *et al*. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **4**, 29–40.

De Hoon,M.J.L. *et al*. (2004) Predicting gene regulation by sigma factors in *Bacillus subtilis* from genome-wide data. *Bioinformatics*, **20**, i101–i108.

Friedman,N. *et al*. (2000) Using Bayesian network to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Ge,H. *et al*. (2002) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerivisiae*. *Nat. Genet.*, **29**, 482–486.

Giaever,G. *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.

Hartemink,A.J. *et al*. (2002) Combining location and expression data for principled discovery of genetic regulatory network models. *Pac. Symp. Biocomput.*, **7**, 437–449.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T. *et al*. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

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

Imoto,S. *et al*. (2003) Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. *J. Bioinformatics Comput. Biol.*, **1**, 231–252.

Imoto,S. *et al*. (2004) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *J. Bioinformatics Comput. Biol.*, **2**, 77–98.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jansen,R. *et al*. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Konishi,S. *et al*. (2004) Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27–43.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Manning,B.D. *et al*. (1999) Differential regulation of the Kar3p kinesin-related protein by two associated proteins, Cik1p and Vik1p. *J. Cell Biol.*, **144**, 1219–1233.

McMillan,J.N. *et al*. (1999) The morphogenesis checkpoint in *Saccharomyces cerevisiae*: cell cycle control of Swe1p degradation by Hsl1p and Hsl7p. *Mol. Cell. Biol.*, **19**, 6929–6939.

Mewes,H.W. *et al*. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Nariai,N. *et al*. (2004) Using protein–protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pac. Symp. Biocomput.*, **9**, 336–347.

Pe'er,D. *et al*. (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**, S215–S224.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Segal,E. *et al*. (2003a) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Segal,E. *et al*. (2003b) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **19**, i264–i272.

Segal,E. *et al*. (2003c) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19**, i273–i282.

Shmulevich,I. *et al*. (2002) Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, **18**, 261–274.

Shou,W. and Deshaies,R.J. (2002) Multiple telophase arrest bypassed (tab) mutants alleviate the essential requirement for Cdc15 in exit from mitosis in *S. cerevisiae*. *BMC Mol. Biol.*, **3**, 4.

Spellman,P. *et al*. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tamada,Y. *et al*. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, ii227–ii236.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.