Using Protein-Protein Interactions for Refining Gene Networks Estimated from Microarray Data by Bayesian Networks

N. Nariai, S. Kim, S. Imoto, and S. Miyano

Pacific Symposium on Biocomputing 9:336-347(2004)

USING PROTEIN-PROTEIN INTERACTIONS FOR REFINING GENE NETWORKS ESTIMATED FROM MICROARRAY DATA BY BAYESIAN NETWORKS

N. NARIAI, S. KIM, S. IMOTO, S. MIYANO
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan

We propose a statistical method to estimate gene networks from DNA microarray data and protein-protein interactions. Because physical interactions between proteins or multiprotein complexes are likely to regulate biological processes, using only mRNA expression data is not sufficient for estimating a gene network accurately. Our method adds knowledge about protein-protein interactions to the estimation method of gene networks under a Bayesian statistical framework. In the estimated gene network, a protein complex is modeled as a virtual node based on principal component analysis. We show the effectiveness of the proposed method through the analysis of *Saccharomyces cerevisiae* cell cycle data. The proposed method improves the accuracy of the estimated gene networks, and successfully identifies some biological facts.

1 Introduction

The complete DNA sequences of many organisms, such as yeast, mouse, and human, have recently become available. Genome sequences specify the gene expressions that produce proteins of living cells, but how the biological system as a whole really works is still unknown. Currently, a large number of gene expression data and protein-protein (p-p) interaction data have been collected from high-throughput analyses, and estimating gene networks from these data has become an important topic in systems biology.

Several methods have been proposed for estimating gene networks from microarray data by using Boolean networks[1,30], differential equation models[3,7], and Bayesian networks[8,9,12,13,14,15,16,22]. However, using only microarray data is not sufficient for estimating gene networks accurately, because the information contained in microarray data is limited by the number of arrays, their quality, noise and experimental errors. Therefore, the use of other biological knowledge together with microarray data is a key for extracting more reliable information. Hartemink *et al.*[13] noticed this idea previously and proposed a method to use localization data combined with microarray data for estimating a gene network. There are other works combining microarray data with biological knowledge, such as DNA sequences of promoter elements[23,32] and transcriptional bindings of regulators[26,27,29].

In this paper, we propose a statistical method for estimating gene networks from microarray data and p-p interactions by using a Bayesian network model. We extract 9,030 physical interactions from the MIPS database[21] to add knowledge about p-p interactions to the estimation method of gene networks. If multiple genes will form a protein complex, then it is natural to treat them as one variable in the estimated gene network. In addition, in the estimated gene network, a protein complex is modeled as a virtual node based on principal component analysis. That is, the protein complexes are dynamically found and modeled based on the proposed method while we estimate a gene network.

Previously, Segal *et al.*[28] proposed a method for identifying pathways from microarray data and p-p interaction data. A different point of our method is that we model protein complexes directly in the Bayesian network model aimed at refining the estimated gene network. Also, our method can decide whether we make a protein complex based on our criterion.

We evaluate our method through the analysis of *Saccharomyces cerevisiae* cell cycle gene expression data[31]. First, we estimated three gene networks, by microarray data alone, by p-p interactions alone, and by our method. Then, we compared them with the gene network compiled by KEGG for evaluation. We successfully show that the accuracy of the estimated gene network is improved by our approach. Second, among 350 cell cycle related genes, we found 34 gene pairs as protein complexes. In reality, most of them are likely to form protein complexes considering biological databases and existing literature. Third, we show an example to use an additional information "phase" together with the microarray data and p-p interactions for estimating a more meaningful gene network.

2 Bayesian Network Model with Protein Complex

Bayesian networks (BNs) are a type of graphical model that represents relationships between variables. That is, for each variable there is a probability distribution function whose definition depends on the edges leading into the variable. A BN is a directed acyclic graph (DAG) encoding the Markov assumption that each variable is independent of its non-descendants, given just its parents. In the context of BNs, a gene is regarded as a random variable and shown as a node in the graph, and a relationship between the gene and its parents is represented by the conditional probability. Thus, the joint probability of all genes can be decomposed as the product of the conditional probabilities.

Suppose that we have n set of microarray data $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ of p genes. A BN model is then written as $f(x_{i1}, ..., x_{ip}|\boldsymbol{\theta}_G) = \prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij}, \boldsymbol{\theta}_j)$, where $\boldsymbol{p}_{ij}$ is the parent observation vector of jth gene (gene$_j$) measured by ith array. For example, if gene$_2$ and gene$_3$ are parents of gene$_1$, we set $\boldsymbol{p}_{i1} = (x_{i2}, x_{i3})^T$. If we ignore the information of p-p interactions, the relationship between x_{ij} and $\boldsymbol{p}_{ij}$ can be modeled by using a nonparametric additive regression model[14,16]

$$x_{ij} = \sum_k m_{jk}(p_{ik}^{(j)}) + \varepsilon_{ij}, \qquad i = 1, ..., n;\ j = 1, ..., p, \tag{1}$$

where $p_{ik}^{(j)}$ is the kth element of $\boldsymbol{p}_{ij}$, m_j is a regression function and ε_{ij} is a random variable with a normal distribution with mean 0 and variance σ_j^2.

When a gene is regulated by a protein complex, it is natural that we consider a protein complex as a direct parent. Therefore, we consider the use of virtual nodes corresponding to protein complexes in the BN model. Concretely, if gene$_2$ and gene$_3$ make a protein complex and regulate gene$_1$, we construct a new variable "complex$_{23}$" from the expression data of gene$_2$ and gene$_3$. In the BN model, then, we consider the relation "complex$_{23}$ → gene$_1$" instead of "gene$_2$ → gene$_1$ ← gene$_3$".

If genes make a protein complex, it is expected that there may be a relatively high correlation among the expression values of those genes. For constructing a new variable representing a protein complex, therefore, we use principal component analysis[17] (PCA). By using PCA, we can reduce the dimension of the data with the least loss of information. Suppose that genes from gene$_1$ to gene$_d$ make a protein complex and that the d dimensional vector $\boldsymbol{a}_1^{[1-d]}$ is the first eigenvector of the matrix $\boldsymbol{S}^{[1-d]} = \sum_i (\boldsymbol{x}_i^{[1-d]} - \bar{\boldsymbol{x}}^{[1-d]})(\boldsymbol{x}_i^{[1-d]} - \bar{\boldsymbol{x}}^{[1-d]})^T/n$ with $\boldsymbol{x}_i^{[1-d]} = (x_{i1}, ..., x_{id})^T$ and $\bar{\boldsymbol{x}}^{[1-d]} = \sum_i \boldsymbol{x}_i^{[1-d]}/n$. Here $\boldsymbol{x}^T$ is the transpose of $\boldsymbol{x}$. The ith observation of the protein complex is then obtained by $c_i^{[1-d]} = \boldsymbol{a}_1^{[1-d]T}(\boldsymbol{x}_i^{[1-d]} - \bar{\boldsymbol{x}}^{[1-d]})$. In such case, we use the regression function $m_{j,[1-d]}(c_i^{[1-d]})$ instead of the additive regression function $m_{j1}(x_{i1}) + \cdots + m_{jd}(x_{id})$. Figure 1 shows an example of modeling a protein complex. *SPC97* and *SPC98* form a protein complex. The solid line is the first principal component and the observations of the protein complex are obtained by projecting expression data onto this line.

This model can be viewed as an extension of principal component regression[2], in which we choose whether we make protein complexes based on our criterion that evaluates the goodness of the BN model as a gene network.

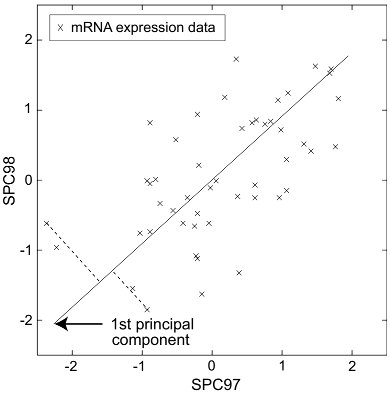

Figure 1: An example of modeling a protein complex by using principal component analysis. The scatter plot of *SPC97* and *SPC98*, and the first principal component are shown.

3 Criterion and Algorithm for Estimating a Gene Network

From a Bayesian statistical viewpoint, we can choose the graph structure by maximizing the posterior probability of the graph G

$$\pi(G|\boldsymbol{X}) \propto \pi(G) \int \prod_{i=1}^{n} f(x_{i1}, ..., x_{ip}|\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G, \tag{2}$$

where $\pi(G)$ is a prior probability of the graph G, $\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ is the prior distribution on the parameter $\boldsymbol{\theta}_G$ and $\boldsymbol{\lambda}$ is the hyperparameter vector. The marginal likelihood measures the closeness between microarray data and the graph G. We add the knowledge about p-p interaction into $\pi(G)$. Following the result of Imoto *et al.*[15], we can model the knowledge about p-p interaction as a prior probability of graph G by using the Gibbs distribution[10].

Let U_{ij} be the interaction energy of the edge from gene$_i$ to gene$_j$ and categorized into 2 values, H_1 and H_2 ($H_1 < H_2$). If there is a p-p interaction between gene$_i$ and gene$_j$, we set $U_{ij} = U_{ji} = H_1$. The total energy of the graph G can then be defined as $E(G) = \sum_{\{i,j\} \in G} U_{ij}$, where the sum is taken over the existing edges in the graph G. The probability $\pi(G)$ is naturally modeled by the Gibbs distribution of the form $\pi(G) = Z^{-1} \exp\{-\zeta E(G)\}$, where ζ ($>$ 0) is an inverse temperature and Z is the partition function given by $Z = \sum_{G \in \mathcal{G}} \exp\{-\zeta E(G)\}$. Here $\mathcal{G}$ is the set of possible graphs. By replacing ζH_1 and ζH_2 with ζ_1 and ζ_2, respectively, the prior probability $\pi(G)$ is specified by ζ_1 and ζ_2. Hence, we have $\pi(G) = Z^{-1} \prod_{\{i,j\} \in G} \exp(-\zeta_{\alpha(i,j)})$, with $\alpha(i,j) = k$

for $U_{ij} = H_k$.

For computing the marginal likelihood represented by the integration in (2), we used the Laplace approximation for integrals[6,19,33] and the result was shown by Imoto *et al.*[14]. Hence, we have a Bayesian information criterion, named BNRC (Bayesian network and Nonparameteric Regression Criterion), for evaluating networks

$$\mathrm{BNRC}(G) = 2\log Z + 2\sum_{\{i,j\}\in G} \zeta_{\alpha(i,j)} + \log\left|\frac{n}{2\pi}J_\lambda(\hat{\boldsymbol{\theta}}_G)\right| - 2nl_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X}), \qquad (3)$$

where

$$\begin{aligned} l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}) &= \frac{1}{n}\sum_{i=1}^{n}\log f(x_{i1},...,x_{ip}|\boldsymbol{\theta}_G) + \frac{1}{n}\log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}), \\ J_\lambda(\boldsymbol{\theta}_G) &= -\partial^2\{l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})\}/\partial\boldsymbol{\theta}_G\partial\boldsymbol{\theta}_G^T \end{aligned}$$

and $\hat{\boldsymbol{\theta}}_G$ is the mode of $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X})$. We can choose the graph structure as the minimizer of BNRC.

Based on the BN model with protein complex and the information criterion described above, we can naturally obtain the greedy hill-climbing algorithm for finding and modeling protein complexes and estimating a gene network as follows:

Step1. For gene$_i$, perform one of four procedures, "add a parent", "remove a parent", "reverse the parent-child relationship" and "none", which gives the lowest BNRC score. If directed cycles are formed, we cancel the operation.

Step2. In Step1, if "add a parent" was performed, go to Step3. Otherwise, go to Step6.

Step3. If the relation between gene$_i$ and the added gene (we denote gene$_{(i)}$) is listed in p-p interactions, go to Step4. Otherwise, go to Step6.

Step4. Construct a protein complex from the expression values of gene$_i$ and gene$_{(i)}$ based on the principal component analysis.

Step5. If the protein complex works better than only using gene$_i$ or gene$_{(i)}$ as a parent of each child of gene$_i$ or gene$_{(i)}$, we use this protein complex in the estimated network. Otherwise, we ignore this protein complex.

Step6. If the BNRC score becomes unchanged, the learning is finished. Otherwise, go to Step1 and continue the greedy hill-climbing algorithm.

Table 1: Comparison result of the cell cycle pathway in KEGG. "agree", "reverse", "false negative" and "false positive" edges are counted by comparing the estimated networks with the KEGG pathway. Note that edges among protein complexes are not counted in this table.

edge type	*using only microarray data*	*using only p-p interactions*	*our method*
agree	4	19	16
reverse	2	(directions unknown)	4
false negative	20	26	18
false positive	55	11	14

4 Computational Experiments

We apply our method to *Saccharomyces cerevisiae* cell cycle microarray data[31], and 9,030 p-p interaction data extracted from MIPS database[21]. For the prior probability $\pi(G)$ given in Section 3, we choose 0.5 for ζ_1 and 25.0 for ζ_2 experimentally. This point is where the maximum number of protein complexes is observed in the estimated gene networks. When we use a larger ζ_1 and a smaller ζ_2, p-p interactions did not contribute to the gene network refinement. On the other hand, when we used a smaller ζ_1 and a larger ζ_2, the resulting network reflected the p-p interactions too strongly.

4.1 Cell Cycle Pathway in KEGG

For evaluating the accuracy of estimated gene networks, we choose 99 genes from KEGG pathway database of *Saccharomyces cerevisiae* cell cycle[18]. In this analysis, we focus on how the accuracy of the estimated network increases by adding the information of p-p interactions. We estimated three gene networks, by using only microarray data, by using only p-p interactions, and by using the proposed method. Then, we compared them with the gene network compiled by KEGG for evaluation.

Table 1 summarizes the result of the comparison among three networks. Note that in this table, edges among protein complexes are not counted, because these edges should not be considered as "gene regulation" in the gene network. By comparing the network estimated by microarray data alone with the network estimated by our method, we can immediately find that the number of edges that agree with KEGG pathway, denoted as *agree*, adequately increases by adding p-p interactions to microarray data. We can also observe that the proposed method can reduce the *false positive* edges drastically. By comparing the network estimated by p-p interactions alone with the network

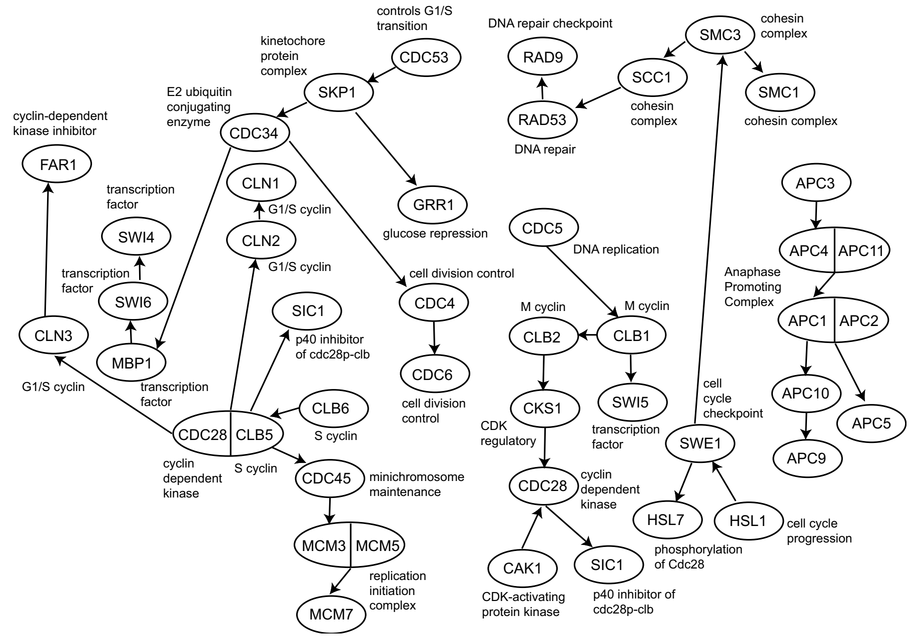

Figure 2: Cell cycle gene network estimated by our method.

estimated by our method, we can find that several *false negative* edges of p-p interactions are newly estimated by adding microarray data, though the number of *agree* edges is almost the same. As for *false positive* edges, we could not observe apparent improvements by adding microarray data.

Figure 2 shows a part of the estimated gene network based on the proposed method. We can find that the proposed method succeeded in finding APC (Anaphase Promoting Complex), MCM (Mini-Chromosome Maintenance) complex, and clb5-cdc28p complex.

4.2 Gene Network with 350 Cell Cycle Genes

For evaluating our method in the sense of modeling a protein complex, we chose 350 genes from the MIPS functional category “mitotic cell cycle and cell cycle control”, and searched protein complexes while learning gene networks. We found 34 candidate protein complexes listed in Table 2.

Among 34 candidate protein complexes, 22 pairs are also listed in the MIPS complex catalogue, and six pairs are reported in existing literature.

Table 2: Detected protein complexes among 350 cell cycle genes. The word *rate* means the contribution rate of the 1st principal component of two genes, and *eval.* means the evaluation of the results. "○" shows that the MIPS protein complexes catalogue contains the pair as a protein complex. "△" shows that while the MIPS catalogue does not contain those pairs, existing literature supports them. "?" shows that the result has not been reported yet.

gene A	*gene B*	*rate*	*eval.*	*annotation*
RSC6	RSC8	0.91	○	RSC complex
MCM5	MCM7	0.89	○	MCM complex
SPC97	SPC98	0.80	○	gamma-tubulin complex
CIK1	KAR3	0.70	○	kinesin-related motor proteins
CLB5	CDC28	0.69	○	clb5-cdc28p complex
GIM3	PAC10	0.67	○	gim complex
SKP1	CDC53	0.66	○	SCF complex
CDC11	CDC12	0.80	○	septin filaments
CDC3	SHS1	0.55	○	septin filaments
CDC10	SHS1	0.54	○	septin filaments
APC1	APC10	0.75	○	APC complex
APC4	CDC23	0.74	○	APC complex
APC4	APC11	0.73	○	APC complex
APC10	APC11	0.72	○	APC complex
APC9	APC10	0.71	○	APC complex
APC1	CDC23	0.66	○	APC complex
APC2	CDC16	0.66	○	APC complex
APC9	CDC16	0.66	○	APC complex
APC1	CDC26	0.64	○	APC complex
APC2	APC5	0.63	○	APC complex
APC3	CDC16	0.63	○	APC complex
APC11	CDC26	0.55	○	APC complex
SMC1	SMC3	0.84	△	cohesin complex[11]
SCC3	SMC3	0.63	△	cohesin complex[11]
BIM1	TUB1	0.69	△	tublin complex[25]
CLN2	CDC53	0.64	△	G1/S transition[34]
CKS1	CDC28	0.57	△	cyclin-dependent kinase[24]
HSL7	SWE1	0.55	△	septin assembly checkpoint[5]
RAD23	RPT6	0.82	?	proteasome
NUF2	NUM1	0.80	?	nuclear migration
NUF1	SPC97	0.79	?	nuclear migration
NUF2	SMC1	0.77	?	nuclear migration
CBF2	YGR179C	0.65	?	centromere/kinetochore-associated
CDC24	SWE1	0.55	?	serine/threonine protein kinase

Although six pairs, denoted as "?" in Table 2, are unknown, they may suggest that each pair forms a protein complex. For example, *RAD23* and *RPT6* may form a protein complex that involves in proteasome activity. In a similar way, *NUF*2 and *NUM1* may work together for nuclear migration. There are 309 p-p interactions among 350 cell cycle related genes, in which only 119 interactions are in fact protein complex related. These results suggest that our method successfully models the protein complexes, and finds the biologically plausible protein complexes.

4.3 Using Phase Information together with Microarrays and P-P Interactions

In this section, we show a case to use an additional information "phase" together with the microarray data and p-p interactions. It is known that cyclins "*CLN1* and *CLN2*", "*CLB5* and *CLB6*", and "*CLB1* and *CLB2*" are activated in G1/S, S, and M phases, respectively[4]. Before estimating a gene network, we choose phase-specific genes whose expression levels are highly correlated with each cyclin listed above. We collected 33 genes from the correlations, i.e., the correlation is greater than 0.75. Also, we selected 93 genes that show p-p interactions with 33 genes and six cyclins. That is, in this analysis, we focus on the gene network with 132 genes. Figure 3 shows the expression patterns of genes that are divided into three groups by the correlations and p-p interactions.

At first, we estimate a gene network for each phase, i.e., G1/S, S and M phases. We then combine those three networks and obtain a final network shown in Figure 4. Genes that are on the dotted line are selected as a member of both phases, i.e., *YOX1* belongs to G1/S phase and also S phase. In this analysis, we can find biologically important genes, such as *HCM1*, *FKH2* and *ACE2*. These genes are transcription factors[20,35], and *FKH2* was reported[36] as a regulator of *CLB2*, *SWI5*, and *HST3*. Although KEGG pathway does not include those genes, we succeeded in finding those important relationships based on our approach.

5 Discussion

In this paper we proposed a statistical method for estimating gene networks by combining microarray gene expression data and p-p interactions. We also proposed a method for modeling protein complexes in the estimated gene network by using principal component analysis. An advantage of our method is that not only p-p interactions, but also protein complexes are naturally modeled under a Bayesian statistical framework. By adding p-p interaction data into our Bayesian network estimation method, we successfully estimated the gene

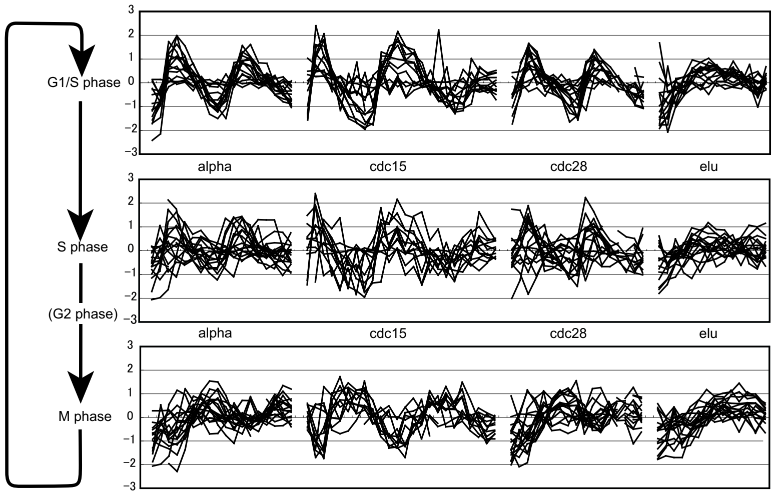

Figure 3: Gene expression profiles that belong to (Top) G1/S phase, (Middle) S phase, and (Bottom) M phase.

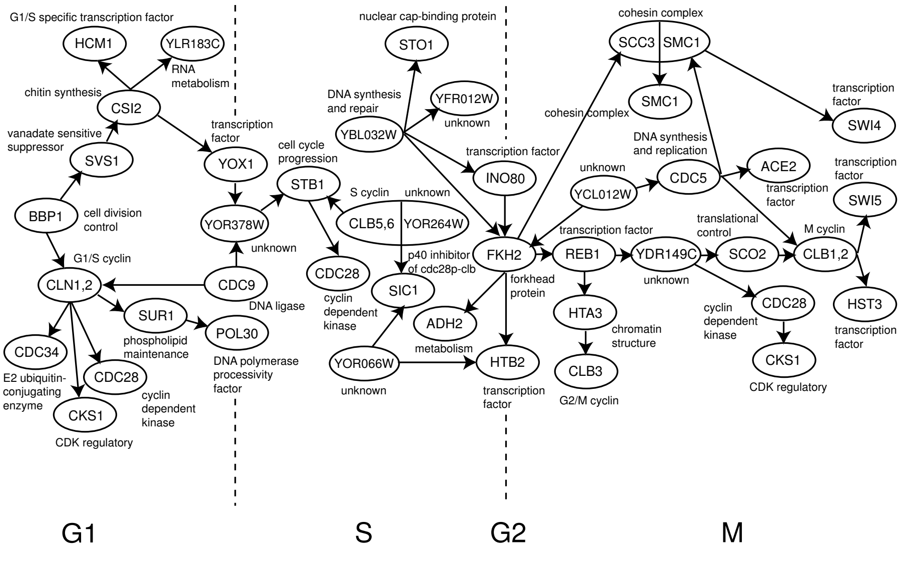

Figure 4: Cell cycle gene network estimated by using "phase" information together with microarray data and p-p interactions.

network more accurately than using only microarray data. We also observed that protein complexes were correctly found and modeled while learning gene networks.

We consider the following topics as our future works: First, currently our greedy algorithm only merges protein pairs based on PCA. Modeling a larger protein complex in the gene network will be an important problem. Second, as real biological processes are often condition specific, it is important to take "conditions" or "environments" into account. Third, in the last experiment, we showed an example that we added an additional information "phase" to the microarray data and p-p interaction data, and estimated a gene network based on those three types of data. We expect that estimating an accurate gene network by using further genomic data, including DNA-protein interactions, binding site information, and so on, will give us more meaningful information about biological processes. We would like to investigate these topics in our future papers.

Acknowledgements

The authors would like to thank three referees for their helpful comments and suggestions.

References

1. T. Akutsu, S. Miyano and S. Kuhara, *Pac. Symp. Biocomput.*, **4**, 17 (1999).
2. S. Chatterjee and B. Price, *John Wiley and Sons*, (1977).
3. T. Chen, H. L. He and G. M. Church, *Pac. Symp. Biocomput.*, **4**, 29 (1999).
4. R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart and R. W. Davis, *Molecular Cell*, **2**, 65 (1998).
5. V. J. Cid, M. J. Shulewitz, K. L. McDonald and J. Thorner, *Mol. Biol. Cell*, **12**, 1645 (2001).
6. A. C. Davison, *Biometrika*, **73**, 323 (1986).
7. M. J. L. de Hoon, S. Imoto, K. Kobayashi, N. Ogasawara and S. Miyano, *Pac. Symp. Biocomput.*, **8**, 17 (2003).
8. N. Friedman, M. Goldszmidt, *in M.I.Jordan ed., Kluwer Academic Publishers*, 421 (1998).
9. N. Friedman, M. Linial, I. Nachman and D. Pe'er, *J. Comp. Biol*, **7**, 601 (2000).
10. S. Geman and D. Geman, *IEEE TPAMI*, **6**, 721, (1984).
11. C. H. Haering, J. Löwe, A. Hochwagen and K. Nasmyth, *Molecular Cell*, **9**, 773 (2002).

12. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola and R. A. Young, *Pac. Symp. Biocomput.*, **6**, 422 (2001).
13. A. J. Hartemink, D. K. Gifford, T. S. Jaakkola and R. A. Young, *Pac. Symp. Biocomput.*, **7**, 437 (2002).
14. S. Imoto, T. Goto and S. Miyano, *Pac. Symp. Biocomput.*, **7**, 175 (2002).
15. S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara and S. Miyano, *Proc. 2nd IEEE Computer Society Bioinformatics Conference*, 104 (2003).
16. S. Imoto, S. Kim, T. Goto, S. Aburatani, K. Tashiro, S. Kuhara and S. Miyano, *J. Bioinformatics and Comp. Biol.*, **1**(2), 231 (2003).
17. I. J. Jolliffe, *Springer-Verlag, New York*, (1986).
18. M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, *Nucleic Acids Res.*, **30**, 42 (2002).
19. S. Konishi, T. Ando and S. Imoto, *Biometrika*, (2003) in press.
20. H. J. McBride, Y. Yu and D. J. Stillman, *J. Biol. Chem*, **274**, 21029 (1999).
21. H. W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkoetter, S. Rudd and B. Weil, *Nucleic Acids Res.*, **30**(1), 31 (2002).
22. D. Pe'er, A. Regev, G. Elidan and N. Friedman, *Bioinformatics*, **17**, S1 (2001).
23. Y. Pilpel, P. Sudarsanam and G. M. Church, *Nature Genetics*, **29**, 153 (2001).
24. G. J. Reynard, W. Reynolds, R. Verma and R. J. Deshaies, *Mol. Cell. Biol.*, **20**, 5858 (2000).
25. K. Schwartz, K. Richards and D. Botstein, *Mol. Biol. Cell*, **8**, 2677 (1997).
26. E. Segal, Y. Barash, I. Simon, N. Friedman and D. Koller, *RECOMB*, 273 (2002).
27. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, *Nature Genetics*, **34**(2), 166 (2003).
28. E. Segal, H. Wang and D. Koller, *Bioinformatics*, **19**, S264 (ISMB 2003).
29. E. Segal, R. Yelensky and D. Koller, *Bioinformatics*, **19**, S273 (ISMB 2003).
30. I. Shmulevich, E. R. Dougherty, S. Kim and W. Zhang, *Bioinformatics*, **18**, 261 (2002).
31. P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein and B. Futcher, *Mol. Biol. Cell*, **9**, 3273 (1998).
32. Y. Tamada, S. Kim, H. Bannai, S. Imoto, K. Tashiro, S. Kuhara and S. Miyano, *Bioinformatics*, (ECCB 2003). in press.
33. L. Tinerey and J. B. Kadane, *J. Amer. Statist. Assoc.*, **81**, 82 (1986).
34. A. R. Willems, S. Lanker, E. E. Patton, K. L. Craig, T. F. Nason, N. Mathias, R. Kobayashi, C. Wittenberg and M. Tyers, *Cell*, **86**, 453 (1996).
35. G. Zhu and T. N. Davis, *Biochim. Biophys. Acta.*, **1448**(2), 236 (1998).
36. G. Zhu, P. T. Spellman, T. Volpe, P. O. Brown, D. Botstein, T. N. Davis and B. Futcher, *Nature*, **406**, 90 (2000).