Modeling Gene Networks Utilizing Evolutionary Information Using Bayesian Network Models

Yoshinori Tamada[1] tamada@kuicr.kyoto-u.ac.jp

Hideo Bannai[2] bannai@ims.u-tokyo.ac.jp

Seiya Imoto[2] imoto@ims.u-tokyo.ac.jp

Toshiaki Katayama[2] ktym@hgc.jp

Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

Satoru Miyano[2] miyano@ims.u-tokyo.ac.jp

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

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

Keywords: gene network, microarray data, evolutionary information

1 Introduction

Estimating the regulatory relationships between genes from genomic data such as microarray gene expression data [1], genomic sequence data [5], or biological databases [2] is becoming a very important problem in systems biology. Recent studies have revealed that highly conserved proteins that exhibit similar expression patterns in different organisms, have almost the same function in each organism. Such conserved proteins are also known to play similar roles in terms of the regulations of genes. Therefore, these evolutionarily conserved information can be used to refine regulatory relationships among genes, which are estimated from gene expression data. Here we propose a statistical method that estimates gene networks from gene expression data and the information of evolutionarily conserved proteins among distinct organisms using Bayesian network models [2, 4]. Our method utilizes the evolutionary information between two organisms and simultaneously estimates two gene networks so that the network of one organism helps to estimate the other organism's network and *vice versa*.

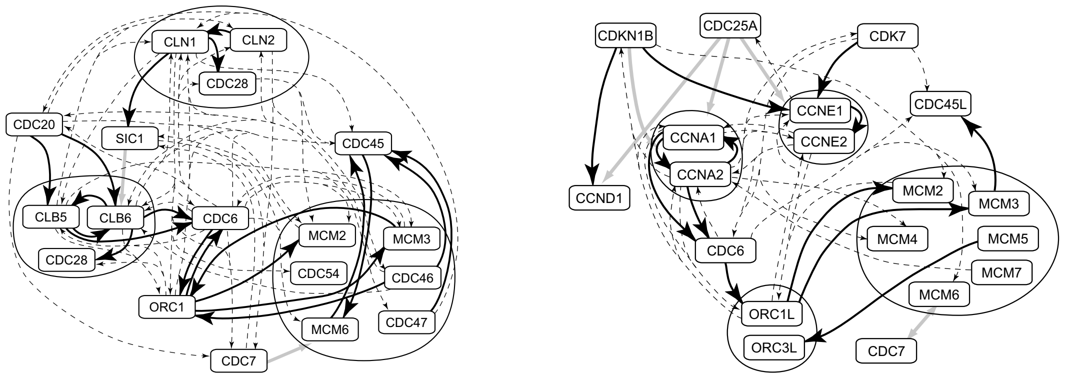
2 Methods

2.1 Bayesian Network Model

A gene network is a graphical model that represents the regulatory relationships between genes. We model a gene network G as a Bayesian network [2, 4], where genes are represented by random variables and the structure is described as a directed graph with the random variables as its nodes. We estimate a gene network from gene expression data $\boldsymbol{X}$ which is a matrix whose element x_{ij} corresponds to the expression value of the jth gene in the ith array, where $i = 1, \dots, n$, $j = 1, \dots, p$. Here, n and p represent the number of microarrays and genes, respectively. In the context of Bayesian networks, the probability of $\boldsymbol{X}$ conditional on G can be decomposed as a product of conditional probabilities, $P(\boldsymbol{X}|G) = \prod_{i=1}^{n} \prod_{j=1}^{p} P(X_{ij}|Pa(X_{ij}))$, where X_{ij} is a random variable corresponding to observation x_{ij}, $Pa(X_{ij})$ is the set of variables of the jth gene's parents at experiment i.

2.2 Probabilistic Framework

We consider two organisms A and B. The evolutionary information $\boldsymbol{H}_{AB}$ between A and B is given as the set of gene pairs of A and B, which is calculated by the E-value of a BLAST search. We consider that the gene pairs included in $\boldsymbol{H}_{AB}$ are orthologous gene pairs among organisms A and B. Assume that we are given $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ which are matrices representing gene expression data for organisms A and B, respectively. Suppose that we want to estimate two gene networks, G_A and G_B of organisms A and B, respectively. Under the framework of Bayesian statistics, we estimate two networks



S. cerevisiae *H. sapiens*

Figure 1: Gene networks estimated by the proposed method. The solid edges are consistent with the KEGG pathways. The dashed edges are not contained in the KEGG pathways. The gray edges are known relationships in the KEGG pathways, but are not estimated. The edges within the MCM complex are omitted.

simultaneously by maximizing the following posterior probability function, $P(G_A, G_B|\boldsymbol{X}_A, \boldsymbol{X}_B, \boldsymbol{H}_{AB})$. Under our assumption, this posterior probability can be decomposed as, $P(G_A, G_B|\boldsymbol{X}_A, \boldsymbol{X}_B, \boldsymbol{H}_{AB})$ $= P(\boldsymbol{X}_A|G_A)P(\boldsymbol{X}_B|G_B)P(G_A, G_B|\boldsymbol{H}_{AB})$. The first two probabilities are calculated by the Bayesian network model, and we define the last probability as $P(G_A, G_B|\boldsymbol{H}_{AB}) = Z^{-1}\ \exp(\zeta_{\mathrm{P}}\ n_{\mathrm{P}} + \zeta_{\mathrm{N}}\ n_{\mathrm{N}})$, where Z represents a normalizing constant, n_{P} and n_{N} the number of commonly connected and unconnected edges in G_A and G_B respectively, and ζ_{P} and ζ_{N} the hyperparameters.

3 Results

In order to evaluate the effectiveness of our proposed method, we focus on the cell cycle gene networks of *Saccharomyces cerevisiae* and *Homo sapiens*, and employ the dynamic Bayesian network model with nonparametric regression [4] as the Bayesian network model, since cell cycle time-course gene expression data of *S. cerevisiae* and *H. sapiens* (Spellman *et al.*, 1998; Whitfield *et al.*, 2002) are publicly available. Since *H. sapiens* is a much more complicated organism than *S. cerevisiae*, it is more difficult to estimate regulatory relations between genes of *H. sapiens* than *S. cerevisiae* from only the information on gene expression data. Through the comparison with the KEGG [3] cell cycle pathways and annotations from the gene ontology hierarchy (Ashburner *et al.*, 2000), we have confirmed that our method succeeded in estimating these gene networks more accurately than the previous method, which uses only gene expression data. Furthermore, our method also succeeded in estimating highly possible and unknown relationships which are likely to be novel (Fig. 1).

References

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

[2] Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S., Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks, *J. Bioinform. Comput. Biol.*, 2:77–98, 2004.

[3] Kanehisa, M. and Goto, S., KEGG: Kyoto encyclopedia of genes and genomes, *Nucl. Acids Res.*, 28:27–30, 2000.

[4] Kim, S., Imoto, S., and Miyano, S., Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems*, 75:57–65, 2004.

[5] Tamada, Y., Kim, S., Bannai, H., Imoto, S., Tashiro, K., Kuhara, S., and Miyano, S., Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection, *Bioinformatics*, 19:ii227–ii236, 2003.