Combining Gene Expression Data with DNA Sequence Information for Estimating Gene Networks Using Bayesian Network Model

Yoshinori Tamada1  SunYong Kim2  Hideo Bannai2
tamada@kuicr.kyoto-u.ac.jp  sunk@ims.u-tokyo.ac.jp  bannai@ims.u-tokyo.ac.jp

Seiya Imoto2  Kousuke Tashiro3
imoto@ims.u-tokyo.ac.jp  ktashiro@grt.kyushu-u.ac.jp

Satoru Kuhara3  Satoru Miyano2
kuhara@grt.kyushu-u.ac.jp  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, the University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
3  Graduate School of Genetic Resource Technology, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

Keywords: microarray gene expression data, gene network estimation, promoter motif detection

1 Introduction
We developed a statistical method for estimating gene networks and detecting promoter elements simultaneously [5]. When estimating a gene network from microarray gene expression data alone, a common problem is that the number of microarrays is limited compared to the number of variables in the network model, making accurate estimation a difficult task. Our method overcomes this problem by combining the microarray gene expression data with the DNA sequence information, into a Bayesian network model. The basic idea of our method is that, if a parent gene is a transcription factor (TF), its children may share a consensus motif in their promoter regions of the DNA sequences. Our method detects consensus motifs based on the structure of the estimated network, then estimates the network again using the result of the motif detection. Our method continues this iteration until the network becomes stable.

2 Method
Figure 1 shows the conceptual view of our method. First, we estimate a gene network from microarray data alone using a Bayesian network model [3]. Based on the structure of the network, we then focus on several genes which are regarded as TF candidates in the estimated network, and define sets of genes that may be co-regulated by each TF candidate. A motif detection method [1] is performed for detecting a consensus motif from each set of possibly co-regulated genes. To reflect the network structure into the motif detection method, we exploit scores attached to each gene. These scores are calculated by a Bayesian network method and represent the likelihood of being a child of a certain parent gene. The motif detection method looks for motif patterns whose appearance in the upstream region is most correlated with the scores of genes.

After the motif detection, we revise the structure of the network based on the motif information. Figure 2 shows an example of such correction of edges. After revising the network, we estimate the network again, this time embedding the information of motif existence into a prior probability of the network [2]. This iterative procedure, the motif detection and the network re-estimation, is repeated until the estimated network does not change any more.
Combining Gene Expression with DNA Sequence for Estimating Gene Networks

3 Result

We conducted Monte Carlo simulations to evaluate our method. At first, we designed an artificial gene network, then generated pseudo microarray data from the network and pseudo DNA upstream sequences for each gene. Table 1 represents the result of Monte Carlo simulations. Although the sensitivity did not change largely, the specificity increased drastically when combining the microarray data with the motif information (34.4% → 54.0%).

Next, we applied our method to S. cerevisiae microarray gene expression data. In the motif detection step, our method chose CHA4 as a TF candidate, and searched motifs from its children and grand-children. Using motif information, we successfully correct the direction of some edges. In addition, we found from several genes in the estimated network, a motif taaac which is known as a motif bound by transcription factor SFF [4]. ACE2, known as a SFF regulated gene, is also located near the network on which we focused. These facts may suggest that CHA4 and SFF have some relationship. Furthermore, from both DNA sequences REB1 and ARG2 which contain taaac, we also found a possible MCM1 binding site, located immediately upstream of taaac, like other MCM1-SFF regulated genes such as ACE2 (Table 2).

References


