An Analysis of Cancer Microarrays in the Pathway Context Using Bayesian Networks

Yohsuke Minowa minowa@kuicr.kyoto-u.ac.jp
Susumu Goto goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

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

Keywords: Bayesian networks, gene expression, gene network, multifactorial disease, cancer cell line

1 Introduction

Genetic analysis of multifactorial diseases caused by complicated interactions of two or more genetic and environmental factors is one of the most important themes in the current genetics. But it is difficult to detect such factors, because each independent factor has small effects on phenotype, which is caused by complicated interactions among multiple factors. Therefore, it is meaningful to analyze molecular pathologies of such diseases in the context of genetic networks that can also represent effects of such factors. In this research, we try to detect candidate genes of such diseases by analyzing quantitative whole-genome gene expression data (DNA microarrays) using the Bayesian network method.

2 Methods

We consider two Bayesian network models (Fig. 1). The first model consists of a continuous node (circle) and a discrete node (square). The continuous node has a vector of gene expression ratios which is assumed to follow the binary mixture Gaussian distribution (Fig. 1a) (e.g. one for cancer, and the other for normal). The dicrete node defines proportions of these binary states. The second model additionally includes relationships between continuous nodes (Fig. 1b). In these models, a discrete node represents extrinsic factors (e.g. mutation, environment factor, or other genes which are not included in this experiment), and our purpose is to estimate the effect of such factors in the network context. The inter-gene relations in the network context are determined as follows.

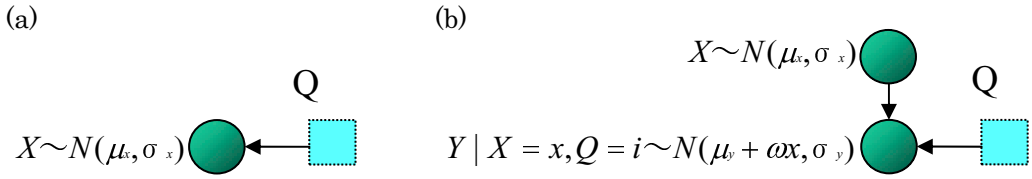

Figure 1: Bayesian network models.

To construct the second model (Fig. 1b), we first defined relationships between genes by adding parent nodes Xs to each node Y according to the mutual information scores I(X,Y) [1]. The number of parent nodes is ranging from 0 to 9. To calculate this score, we first assumed the first model (Fig. 1a) for each node, and performed EM iteration (independently three times, and used average of the parameters) to infer the parameters (μ, σ, proportion of mixture, and probability where each data point belongs to a particular distribution). We can construct probability vectors $P_{N1}$and P_{N2} for each distribution N1 and N2 contained in the binary mixture distribution. Then, we can define I(X,Y) as follow:

$$I(X,Y) = \sum_{i=1}^{2}\sum_{j=1}^{2} P(X \sim N_i, Y \sim N_j) \log \frac{P(X \sim N_i, Y \sim N_j)}{P(X \sim N_i)P(Y \sim Y_j)}$$

where $P(X{\sim}N_i) = p_{1i} + \ldots\ldots + p_{ni}/\ n$ for n data points, p_{ni} is the probability of n'th data point belonging to i'th distribution, $P(Y{\sim}N_j)$ is the same as X, and $P(X{\sim}N_i,\ Y{\sim}N_j) = p_{1i}p_{1j} + \ldots + p_{ni}p_{nj}\ /\ n$.

Next, we assumed the second model (Fig. 1b) for each node using the network structure inferred from mutual information scores, and performed parameter inference using EM algorithm (independently three times, same as before).

We then tested significant difference between the size of mixture of each node, assuming each node has only one status (null hypothesis) or mixture of two status (alternative hypothesis). We rejected null hypothesis at global p-value$<$0.05. In this research, we used NCI60 cancer cell line DNA microarrays [2], which contained 64 cell lines from 9 different tissues (CNS, renal, ovarian, leukaemia, colon, melanoma, breast, prostate, and non-small-lung) with about 8,000 genes. We used the gene set which contained 1,156 genes with at least sevenfold variations relative to the reference in at least 4 cell lines (4/64). We also performed this analysis with partial samples, excluding each cell line sample to determine causation of binary status for each node.

We used the Bayes Net Toolbox software package [3] written by MATLAB to construct Bayesian networks.

3 Results and Discussion

Fig. 2a shows the difference of the number of significant nodes with various values of the number of parent nodes. Significant hits without network context were largely reduced as parent nodes are included. In contrast, some genes are turned out to become significant, when considering the network context (Fig. 2b). Biological significance of these genes will be discussed individually.

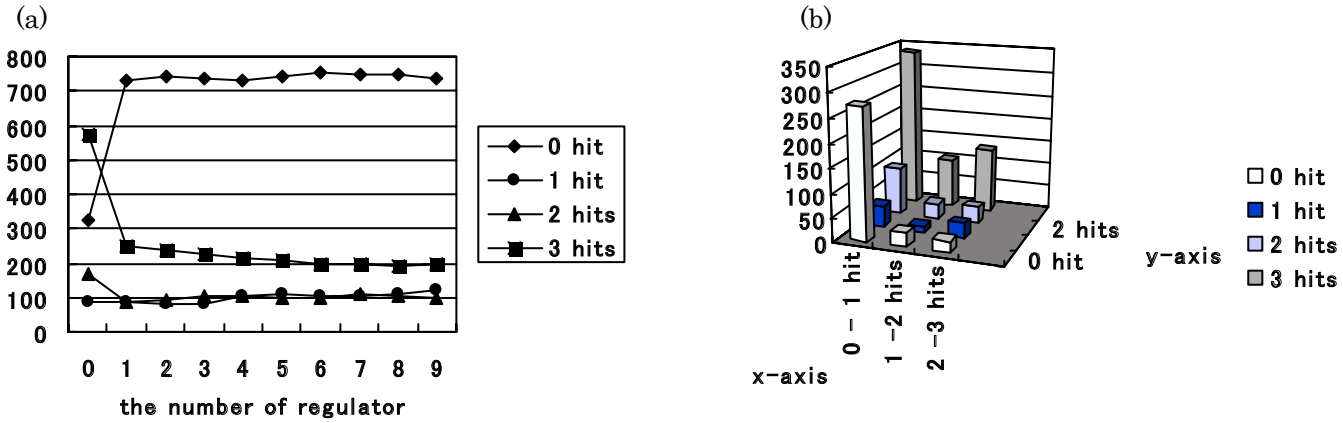

Figure 2: (a) The number of significant nodes plotted against the number of parent nodes (0-9) with the number of hits (0-3) in EM iterations. (b) The number of significant nodes plotted against the numbers of hits with (x-axis) or without (y-axis) network context.

Acknowledgments

This work was partially supported by Grant-in-Aid for JSPS fellows, 14061432, from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Pe'er, D., Regev, A., and Tanay, A., Minreg: Inferring an active regulator set, *Bioinformatics*, 18(Suppl.1):S258–S267, 2002.

[2] Ross, D.T. *et al.*, Systematic variation in gene expression pattern in human cancer cell lines, *Nature Genet.*, 24:27–235, 2000.

[3] `http://www.cs.berkeley.edu/~murphyk/Bayes/bnt.html`