

Operon Prediction by DNA Microarray: An Approach with a Bayesian Network Model

Hitoshi Shimizu

hitosh-s@is.aist-nara.ac.jp

Shigeyuki Oba

shige-o@is.aist-nara.ac.jp

Shin Ishii

ishii@is.aist-nara.ac.jp

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara 630-0192, Japan

Keywords: DNA microarray, operon prediction, Bayesian network

1 Introduction

An operon is a set of genes in prokaryotes, which are transcribed to a single mRNA transcript. Although the operon organization has not yet been completely revealed even in model organisms, such as *E. Coli* and *B. subtilis*, the understanding of the operon organization is important for various transcriptome analyses and for the prediction of genes' functions.

Many methods that predict transcription units use correlation coefficients, denoted by r , between genes, which are calculated from DNA microarray data. Because the r 's distribution of operon pairs (OPs) is different from that of non-operon pairs (NOPs), OPs and NOPs can be discriminated based on r [2]. For example, Bockhorst *et al.* [1] proposed a method with a Bayesian network to link many kinds of observations such as spacer sizes, expression profiles, and codon usage. Tjaden *et al.* [3] used a hidden Markov model (HMM) to predict transcription boundaries.

Although these methods utilize correlations between only adjacent genes on a single DNA strand, a pair of genes that are not immediately next to each other can be an operon pair when the members of the operon are more than two. In this report, we propose a new method for the operon prediction, which utilizes correlations between not only adjacent but also distant genes on a DNA strand.

2 Method

Any pair of genes which are adjacent on a single DNA strand can be divided into two groups, OP (operon pair) or NOP (non-operon pair). The prediction is to divide all pairs of adjacent genes into predicted OPs and predicted NOPs. The prediction is evaluated whether it is consistent with the operon structures determined by biological experiments. Evaluation criteria are sensitivity (correctly predicted OP/known OP) and specificity (correctly predicted NOP/known NOP).

We formulated a Bayesian network model which is assumed to generate correlation of every gene pair (Fig.1 left). $z_{i,j}$ is a hidden variable that takes 0 or 1, corresponding to the case where the pair of gene_{*i*} and gene_{*j*} is NOP or OP, respectively. $r_{i,j}$ is the correlation coefficient between gene_{*i*} and gene_{*j*}, which is assumed to be randomly generated depending on $z_{i,j}$. This generation process, $p(r_{i,j}|z_{i,j})$, is determined beforehand using biologically known operon structures. The problem is to estimate the posterior distribution of $z_{i,j}$ for a given set of $r_{i,j}$ calculated from microarray data. We used variational Bayes method to approximate the posterior probability.

3 Results

In order to evaluate the new method, we applied it to a dataset (correlation coefficient matrix) artificially generated from the known operon structures and microarray data of *B. subtilis*. The ROC

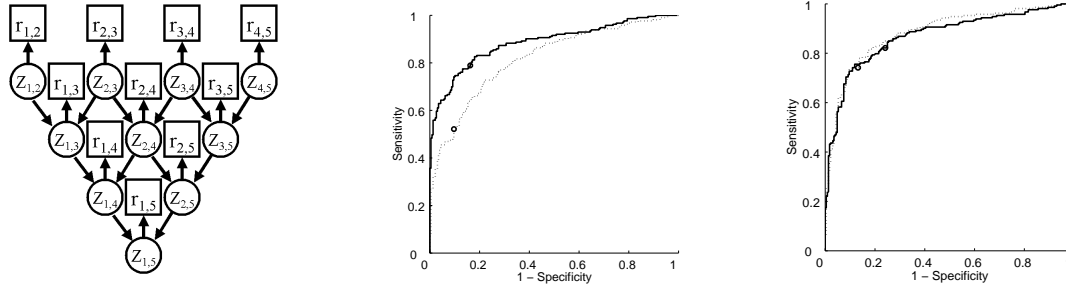


Figure 1 : A Bayesian Network model (left). Squares represent observable variables($r_{i,j}$) and circles represent hidden variables($z_{i,j}$). An arrow stands for dependence. Hidden variables have a hierarchical dependence structure. For example, if a pair of distant genes is OP, any pair of genes between the two genes is always OP. $r_{i,j}$ depends on the corresponding hidden variable, $z_{i,j}$, which means that the distribution of r is dependent on whether the pair is OP or NOP. ROC curve of predictions by our method (line) or the conventional method (dotted line), when applied to an artificial dataset (center). ROC curve of prediction by our method (line) or the conventional method (dotted line), when applied to a *B. subtilis* dataset (right).

curves of prediction by our method and the naive correlation-based method show that the new method is more effective than the naive one (Fig.1 center).

Next, we applied the method to an experimental dataset, and compared the ROC curves of our method and the naive method. The dataset consists of DNA microarray data of *B. subtilis* cultured in various medium conditions (3989 genes \times 88 samples). We used 206 known operons as answers. As a result, there is no significant difference between our method and the naive method (Fig.1 right).

We speculate that the advantage of our method in comparison to the naive method depends on the number of samples. Then, we applied the method to a dataset whose samples were about a half of the original dataset. Although the reduction of samples naturally degraded the performance, the performance of the proposed method was frequently better than that of the naive one (data not shown).

4 Discussion

We showed that our method is effective on the artificial dataset whose generation process was consistent with what we assumed. However, our method did not necessarily show better performance than the naive one when applied to the experimental dataset. A possible reason is that the generation process was apart from the real process due to the short of microarray experiments. In the present study, we determined the generation process of correlation coefficients based on the biologically-known operon structures, but the information amount may not be sufficient.

Acknowledgments

The DNA microarray data were provided by N. Ogasawara and K. Kobayashi (NAIST).

References

- [1] Bockhorst, J., Craven, M., Page, D., Shavlik, J., and Glasner, J., A Bayesian network approach to operon prediction, *Bioinformatics*, 19(10):1227–1235, 2003.
- [2] Sabatti, C., Rohlin, L., Oh, M., and Liao, J.C., Co-expression pattern from DNA microarray experiments as a tool for operon prediction, *Nucleic Acids Research*, 30(13):2886–2893, 2002.
- [3] Tjaden, B., Haynor, D.R., Stolyar, S., Rosenow, C., and Kolker, E., Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis, *Bioinformatics*, 18(S1):S337-S344, 2002.