Quantitative Estimation of Cross-Hybridization in DNA Microarrays Based on a Linear Model

Mitsuteru Nakao nakao@kuicr.kyoto-u.ac.jp
Yoshinori K. Okuji okuji@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

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

Keywords: systematic errors, DNA microarray, cross-hybridization

1 Introduction

Microarray expression data contains possible errors that could be classified into two types, one representing systematic errors and the other representing random errors. Systematic errors may occur due to different factors such as shapes of spots, purity of probes, sequences of reverse transcription (RT) primers, higher order structures of target nucleotides and probe-target cross-hybridization [1, 2]. In this study, we propose a linear model of systematic errors caused by probe-target cross-hybridization. We estimate the influences of such systematic errors so as to reconstruct more realistic gene regulatory networks and co-expression gene groups by the profiles [3].

2 Method and Results

2.1 Data preparation

We used gene expression profiles that were measured by the microarray, CyanoChip [6] by TaKaRa (Kyoto, Japan) that contains almost all ORFs, 3079 ORFs excluding transposases, in the *Synechocystis* sp. PCC6803 genome [4]. The probe DNA sequences were prepared up to the length of 1,000 nt and spotted on the microarray. Gene expression profiles analyzed here were provided by the *Synechocystis* DNA chip consortium.

2.2 Statistical tests

To measure sequence similarities and coverage of the probes and targets, we used `blast` [5] and searched against the genome sequence. As the gene expression similarity metric, we used the Pearson correlation coefficient. In addition, the expression similarities were classified into twenty-one ranks by the correlation coefficient -1.0 to 1.0 with the interval 0.1. All pair-wise gene expression similarities, 4,474,536 pairs, were computed. Their distribution seems to be a bell shape and shifted to a positive correlation. To estimate the influences of cross-hybridizations cased by similar sequences, we divided the probes into two groups, those having similar sequences and those that are unique. We tested these two groups by χ^2 test for independency between the error factor and the expression similarities.

The result shows that the χ^2 statistics dropped to the significance level of 5% when the parameter corresponds the match length of 33 nt and the blast $E - value = 0.01$ for the probe-target pairs. This suggests that the systematic errors would occur in DNA microarray experiments under usual conditions.

2.3 Linear model

The statistical tests show that certain sequence similarities in probes and targets lead to significant influences on observed gene expression similarities. Generally, the probe-target cross-hybridization consists of multiple relations, where j^{th} gene hybridizes with i^{th} probe. The problem can be formulated as follows. A simultaneous equation model describes the cross-hybridization,

$$E_i^O = k_{i,1}E_1^T + \ldots + k_{i,j}E_j^T + \ldots + k_{i,N}E_N^T$$

where E_i^O represents an observed expression level of gene at i^{th} probe, $k_{i,j}$ represents a cross-hybridization coefficient between j^{th} gene product and i^{th} probe and E_i^T represents the true expression level of i^{th} gene product. Considering all N probes and targets, observed expression level $E^O = {}^t(E_1^O, \ldots, E_N^O)$, cross-hybridization coefficient matrix K and the true expression level $E^T = {}^t(E_1^T, \ldots, E_N^T)$, the relation of probe-target cross-hybridization can be expressed as: $E^O = KE^T$ Let K be a directed graph, then a cluster of cross-hybridized genes may be considered as a complete subgraph. Unfortunately E^T is not detectable directly; however, if we can identify K^{-1} that is the inverse matrix of K, E^T would be computable by the equation, $E^T = K^{-1}E^O$. To identify cross-hybridization coefficient matrix K, it would be necessary to perform well designed experiments.

Acknowledgements

We thank all members in the *Synechocystis* DNA chip consortium for *Synechocystis* sp. PCC6803 expression data production and T. Kamiya, S. Asanuma (Kyoto Univ, Japan) and I. Uchiyama (NIBB, Japan) for expression data management.

M.N. was supported by the Research Fellowship of the Japan Society for Promotion of Science for Young Scientists. This work was supported by the Genome Frontier Project of the Science and Technology Agency in Japan. The computational resource was provided by the Supercomputer Laboratory, Kyoto University.

References

[1] Sourthern, E., Mir, K., and Shchepinov, M., Molecular interactions on microarrays, *Nature Genet. supplement* 21:5–9, 1999.

[2] Richmond, S.C., Glasner, D.J., Mau, R., Jin H., and Blattner, R.F., Genome-wide expression profiling in *Escherichia coli* K-12, *Nucleic. Acids Res.* 27:3821–3835, 1999.

[3] Nakao, M., Okuji, K.Y., Itoh, M., Katayama, T., Kawashima, S., Suzuki, I., Murata, N., and Kanehisa, M., Theoretical Estimation of Systematic Errors Caused by Probe-Target Cross-Hybridization in DNA Microarray Experiments., *Proc. JST Symp. Int. Conf. Systems Biology,* 2000 (in press).

[4] Kaneko, T. *et al.,* Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions., *DNA Res.* 3:109–136, 1996.

[5] Altschul, S.F., Gish, W., Miller, E.W. and Lipman, D.J., Basic local alignment search tool., *J. Mol. Biol.* 215:403–410, 1990.

[6] URL http://www.takara.co.jp/