Theoretical Estimation of Systematic Errors Caused by Probe-Target Cross-Hybridization in DNA Microarray Experiments

Mitsuteru Nakao [1]
nakao@kuicr.kyoto-u.ac.jp

OKUJI, K. Yoshinori [1]
okuji@kuicr.kyoto-u.ac.jp

Masumi Itoh [1]
itoh@kuicr.kyoto-u.ac.jp

Toshiaki Katayama [1]
katayama@kuicr.kyoto-u.ac.jp

Shuichi Kawashima [1]
shuichi@kuicr.kyoto-u.ac.jp

Iwane Suzuki [2]
iwane@nibb.ac.jp

Norio Murata [2]
murata@nibb.ac.jp

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

[1] Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
[2] National Institute of Basic Biology, Okazaki, Aichi, 444-8585, Japan
* *To whom correspondence should be addressed*

Abstract

Microarray gene expression data contains many errors that could be classified into two types, one is systematic error and the other is random error. We should estimate the effects of systematic error so as to reconstruct more realistic genetic networks by using gene expression profiles. Systematic errors occur on multiple stages, such as spotting, RT reaction, hybridization and scanning processes. In this study, we have considered especially systematic errors to depend on the probe/target nucleotide sequences, such as the probe-targe sequence similarities and the probes purity.

Statistical tests for the systematic errors have been performed with the *Synechocystis* DNA microarrays. The results show that certain sequence similarities in probes lead to the significant influences in gene expression similarities.

1 Introduction

High-throughput gene expression profiling is one of the most effective approaches to studyregulation programs of the genetic network in the living cell. Attempts have been made to actually reconstruct genetic networks from gene expression profiles, but they usually assume no errors or underestimate possible errors in the profiles when clustering algorithms and learning algorithms are applied [1] [2]. Generally speaking, however, genome-wide gene expression profiling methods, especially microarrays, contain different types of errors on various levels. Systematic errors may occur due to different factors such as shapes of spots, purity of probes, sequences of reverse transcription (RT) primers and higher order structures of target nucleotides. Thus filtering methods to remove odd signals and to minimize these influences have been reported [3].

In addition to systematic errors there are also random errors, but random errors influences can be decreased as more experiments are made. On the other hand, the frequency of systematic errors may still be high regardless of the number of experiments. To decrease such errors, it is necessary to improve the methods of experiments by taking theoretical characteristics of systematic errors into

Type	Reason
Probe/Spot	Probe sequence purity
	Spot shape
	Spot coordination
Target	RT primer (random oligomer or specific primer)
	Primer's Tm
	Frequency of oligonucleotides sequence in mRNAs
	Ensemble of higher structures conformed by mRNAs
Hybridization	Probe-target cross hybridization
	Efficiency of hybridization
Dye	Fluorecent specificity (Cy3, Cy5)
Scanning	Scanner specificity

Table 1: Systematic errors in microarray experiments

consideration. As for systematic errors on DNA microarray experiments, there are probe-target cross-hybridizations caused by the sequence similairties among multiple probes that are spotted on the microarray and multiple targets. Gene expression detection methods, which are based on DNA-DNA hybridizations, have also error factors. For example, probe-target cross-hybridizations at perfectly matching oligonucleotide regions, influences of higher order structures at matching regions, and the differences in the number of RT primer binding sites in the RT reaction. It is known that the differentiation of individual genes in diverse protein families cannot be detected in expression experiments because high sequence similarities in the nucleotide level cause cross-hybridization of multiple probes and multiple targets. Table 1 summarizes different types of systematic errors caused by different steps of microarray experiments.

In this study, we estimated influences of such systematic errors in DNA microarray experiments from a correlation analysis between the error factors and the expression similarities.

2 Data and Methods

2.1 Data preparation

We used gene expression profiles which were measured by the microarray, CyanoChip by TaKaRa(Kyoto, Japan) that contain almost all ORFs in the Synechocystis genome [4]. The probe DNA sequences were aligned up to the length of 1,000 nt in the microarray. Gene expression profiles were provided by the *Synechocystis* DNA chip consortium. The profiles included fifty six experiments, such as for measuring cold-shock, salt, and light stress responses.

The sequence similarities of the probes and targets on the array were searched against *Synechocystis* sp. PCC6803 genome[6] in KEGG/GENES[7] genome database with the probe sequences. *Synechocystis* sp. PCC6803 has 3.5Mb single circular genome and contains 3,166 protein coding ORFs, 6 rRNAs and 43 tRNAs. We used `blastn`[8] to search the sequence similarities.

2.2 Expression similarity

As the gene expression similarity metric, we use the Pearson correlation coefficient[5], r_{xy} of gene x and y, which is given by,

$$r_{xy} \equiv \frac{v_{xy}}{\sqrt{v_{xx}v_{yy}}}$$

where v_{xy} represents the covariance of gene x and gene y expression ratios and v_{xx} or v_{yy} represents the variance of each gene expression ratios. The expression similarities were classified in twenty one ranks by the correlation coefficient -1.0 to 1.0 with the interval 0.1.

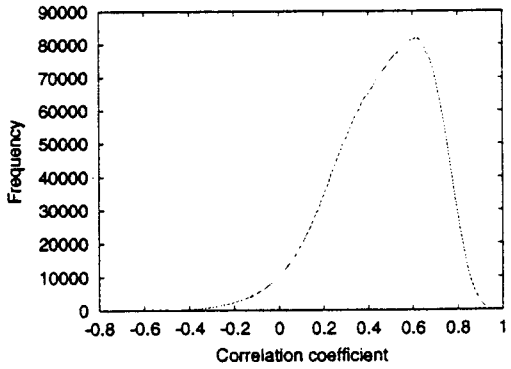

Figure 1: A pairwise distribution of the gene expression similarity in fifty six conditions. The distribution contains 4,474,536 pairs in total and shows the mean as 0.47, the mode as 0.62, the standard deviation as 0.22 and the variance as 0.05.

2.3 Statistical tests

We estimated the influences of errors according to the gene expression similarities. We divided probes into two groups accoding to the parameter of the errors such as probe sequence similarity and probe purity.

χ^2 test for independency between the error factor E and the expression similaritity S is given by,

$$\chi^2(E,S) = \sum_{i=1}^{2}\sum_{j=1}^{21} \frac{(x_{ij} - e_i s_j/N)^2}{e_i s_j/N}$$

where E is divided into two groups, S is divided into twenty one ranks of expression similarity, x_{ij} represents the observation of $i^{th}E$ with $j^{th}S$, $e_i s_j/N$ represents the expected value, e_i represents the number of members that belong to $i^{th}E$, s_j represents the number of members that belong to $j^{th}S$ and N represents the number of the total population.

All data preparations and statistical analysis were investigated by PostgrSQL relational database and Perl scripts.

3 Results and Discussion

3.1 Gene expression similarity

All pairwise gene expression similarities, 4,474,536 pairs, were computed. Pairwise distribution of the expression similarities seems to be shifted to a positive correlation in figure 1. It would suggest that the similar biological systems are shared by many stress response systems in *Synechocystis*. It would also suggest that the broadly diversed protein families and the conserved motifs (cf. ATP binding motif) would occur on the probe-target cross-talk. These effects were estimated in the following section.

3.2 Effect of probe DNA purity

Spotted probes have been processed by the following quality control by TaKaRa. There are three grades of purification on the microarray (personal comunications), the quality being approximately

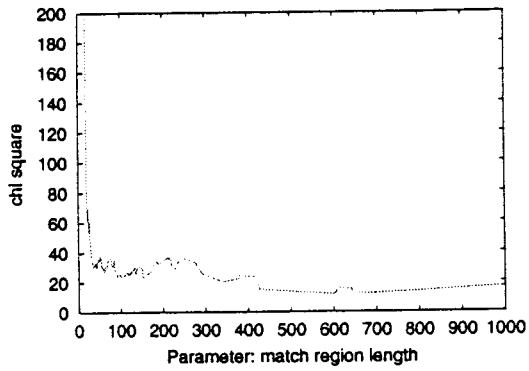

Figure 2: χ^2 test of probe-target nucleotide sequence similarities according to gene expression similarities in fifty six conditions. Critical region is $\chi^2 = 32.7$ on $P = 0.05$ with the degree of freedom of twenty one.

Probe purity	Number of spots	χ^2 statistics	$\chi^2(p = 0.05)$
<70	131	235.6	32.7
>=70	648	3903	32.7
>=90	2316	25744	32.7

Table 2: χ^2 test of the probes purity accoding to gene expression similarities.

"<70 %", ">=70 %" and ">=90 %".

We have tested correlations between the purity of probes and the gene expression similarities to estimate the errors. The result shows that the error caused by the purity of probes was of little significance by the χ^2 test in table 2.

3.3 Effect of probe-target sequence similarity

The results of searching the probe-target sequence similarities show the pairs of probes are impossible to differentiate distinct targets. The similarity was significantly high, the E value of 0.0, the match length of >300 nt and the percent identity of >90% in table 4.

All probe sequences are divided into two groups, whether probe sequences having similarity with other probe targeting cDNAs or not. The similarity search for each probe and ORF in *Synechocystis sp.* PCC 6803 has been investigated.

We tested relationships between the probe-targrt sequence similarities and the gene expression similarities to estimate the errors by the χ^2 test. The result shows that the χ^2 statistics dropped to the significance level of 5 % the parameter = 33 in figure 2. It suggests that the probes can lead to the errors with the 5 % risk level, when probes have sequence similarities with the matching region length of more than 33 and with the E value = 0.01.

	E value <0.1	%identity = 100
Number of similarty pairs	7,416	5,932
Number of ORFs	2,067	1,808
Length of match region	17 .. 1,000	17 .. 211

Table 3: Probe-target sequence similarity pairs within the E value range of <0.1 and the percent identity = 100% by `blastn` in *Synechocystis*.

Probe	Target	Bit	Match	Length	%id	Target Description
sll0986	slr2096	642	367	382	96	hypothetical protein
slr2096	sll0986	642	367	382	96	hypothetical protein
sll1157	slr2096	650	393	417	94	hypothetical protein
slr2096	sll1157	652	393	417	94	hypothetical protein
sll1156	slr1902	664	400	424	94	hypothetical protein
slr1902	sll1156	666	400	424	94	hypothetical protein
slr1902	slr2095	690	403	424	95	hypothetical protein
slr2095	slr1902	690	403	424	95	hypothetical protein
sll1156	slr2095	660	401	426	94	hypothetical protein
slr2095	sll1156	662	401	426	94	hypothetical protein
sll1201	sll1774	918	568	603	94	hypothetical protein
sll1774	sll1201	918	568	603	94	hypothetical protein
sll1774	slr1712	916	573	610	93	hypothetical protein
slr1712	sll1774	916	573	619	93	hypothetical protein
sll1201	slr1712	821	585	642	91	hypothetical protein
slr1712	sll1201	821	585	642	91	hypothetical protein
sll0654	sll0656	1001	610	645	94	nucH; extracellular nuclease
sll0656	sll0654	1001	610	645	94	alkaline phosphatase [EC:3.1.3.1]
slr0927	sll0849	1407	926	998	92	psbD; photosystem II D2 protein
sll0849	slr0927	1411	928	1000	92	psbD2; photosystem II D2 protein
sll1867	slr1311	1957	996	999	99	psbA2; photosystem II D1 protein
slr1311	sll1867	1959	997	1000	99	psbA3, psba-3; photosystem II D1 protein

Table 4: Significant high sequence similarity probe-target pairs with the E-value = 0.0 in the *Synechocystis* genome. Bit, Match, Length and %id mean the `blastn` bit score, the number of matching bases, the length of aligned region, and the percent identity.

4 Conclusions

The statistical tests reported in this study show that certain sequence similarities in probes lead to significant influences on observed gene expression similarities. The error factors pointed out a risk of gene expression analysis by using microarrays, and theoretical considerations of problematic probes would be useful to design new microarray probes and to help reanalyze pervious experiment data. We plan to perform similar analysis of systematic errors in microarray of other organisms, *Bacillus subtilis* and *Saccharomyces cerviecae*.

Acknowledgements

We thank Y. Hihara, A. Kamei, M. Ikeuchi (Univ. Tokyo, Japan) and all other 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. The authors thank K. Ohkubo (Osaka Univ., Japan) for his valuable discussions on errors in DNA-DNA hybridization, and A. Tanaka and S. Goto (Kyoto Univ, Japan) for their comments on this manuscript.

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 resorce was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

References

[1] Brown, P. M., Grundy, W. N., Lin, D., Cristinini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr and Haussler, D., Knowledge-based analysis of microarray gene expression data by using support vector machine, *Proc. Natl. Acad. Sci. USA* **97**, 262-267 (2000).

[2] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M., Systematic determination of genetic network architecture, *Nature Genet.* **22**, 281-285 (1999).

[3] Sourthern, E., Mir, K. and Shchepinov, M., Molecular interactions on microarrays, *Nature Genet. supplement* **21**, 5-9 (1999)

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

[5] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns., *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868 (1998).

[6] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S., 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).

[7] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes., *Nucleic. Acids Res.* **27**, 29-34 (1999).

[8] Altschul, S. F., Gish, W., Miller, E. W. and Lipman, D. J., Basic local alignment search tool., *J. Mol. Biol.* **215**, 403-410 (1990).