

Parametric Treatment of cDNA Microarray Data

Tomokazu Konishi

konishi@agri.akita-pu.ac.jp

Biotechnology Institute, Faculty of Bioresource Sciences, Akita Prefectural University,
Ohgata Minami 2-2, Akita 101-0444, Japan

Keywords: microarray data normalization, data comparison, three-parameter lognormal distribution model

1 Introduction

As each set of microarray data is affected by variations in experimental conditions, appropriate normalization processes are required. Various approaches towards such normalization have been proposed and generally involve adjustments to every pair of the data sets, often between the simultaneously hybridizing R and G probes. Chen *et al.* [1] introduced a model in which each the signal ratio between the probes is normally distributed; the same concept is basically used in the process of the Stanford Microarray Database [6]. Recently, Yang *et al.* [4] extended this method by stabilizing the average of the signal ratio, which could be biased based on signal intensity. Such non-parametric methods have been further improved by introduction of a parameter that also maintains the variance in the signal ratios [2, 3]. An alternative and simple parametric method that assumes lognormal distribution of data has also been widely employed. Besides its simple ease in performing the calculations, it is also capable of determining the data z-scores, a possible common unit for data comparisons. However, microarray data often have a skewed distribution, and this low fidelity to the distribution model severely limits the accuracy of the data.

In the parametric process, the estimation of background, which is defined as the constant part of additive noise [1], can be a major source of normalization inaccuracies. In most cases, the background is estimated from the area outside the DNA spot of the image data; based on the assumption that the background on a tip is uniform. However, as DNA spots and also the intact surface of the tip can bind free dyes at different densities, such estimations are clearly prone to errors. To check this possibility, an alternative estimation method that is based on signal intensity of control DNA is tested against the conventional image-based method. The distributions of both sets of processed data are investigated using probability plots.

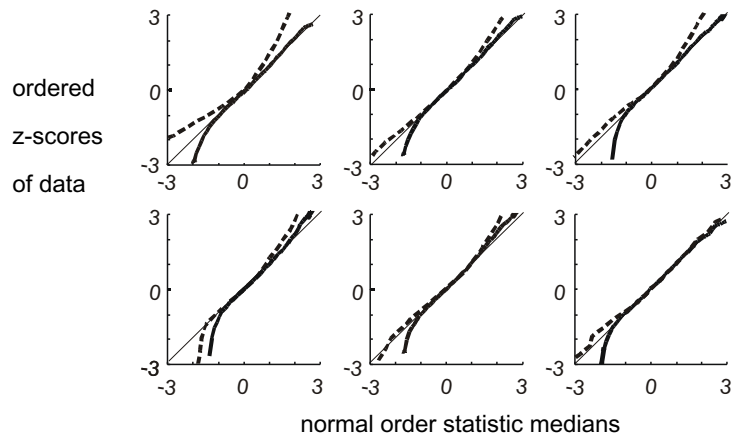
2 Method and Results

Microarray experiments were carried out with three types of DNA tips, containing different sets of rice EST clones [5], that were hybridized with probes derived from rice seedlings. The background was subtracted from the image data by a local estimation program (ArrayVision, Molecular Dynamics) or by signals on control DNA spots that were not expected to hybridize to any of the rice cDNAs. Z-scores of the subtracted data were obtained according to the standard lognormal distribution model [7]. The scale parameter was found as the median and the shape parameter was found as the interquartile range [8] of logarithms of data. Data were processed pin-wise to avoid the systematic errors caused by printing machinery [7].

The parametric nature of the data distribution was confirmed using lognormal probability plots [7]. If data fit the distribution model, the plot would be expected to form a straight line at $y=x$. Such

characteristics were indeed commonly observed over a wide range in data sets in which the background was estimated from control spots (Fig. 1, solid line). In contrast, probability plots of the same data sets using conventional background estimation (dotted line) showed a narrower range of linearity, and thus a poor fit to the model. Over 40 duplicated sets of data collated from 12 different experiments have been analyzed, and all confirm the results presented here.

Figure 1. Examples of probability plots for rice microarray data. The background was estimated by a conventional image-based method (dotted line) or by signal reading on control DNA spots (solid line).



The lowest parts of the plots are always found to bend/skew, a feature commonly observed in signals with z-scores of -1 to -2, suggesting that lowest 16 to 2 % of data, respectively, do not obey the distribution model. Such distortions in the distributions may be due to random additive noise, which can have a more drastic effect on weaker signals. As the magnitudes of the noise variance would be affected by experimental conditions, the breakdown point in each data set also might be affected by such conditions.

3 Discussion

With proper background estimation, a wide range of microarray data is found to obey a lognormal distribution. At least within the applicable range, the data z-scores would be comparable. It appears that image-based background estimation is the main source of inaccuracies in the conventional parametric normalization process.

References

- [1] Chen, Y., Dougherty, E.R., and Bittner, M.L., Ratio-based decisions and the quantitative analysis of cDNA microarray images, *J. Biomed. Optics*, 2:364–374, 1997.
- [2] Durbin, B., Hardin, J., Hawkins, D., and Rocke, D., A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics*, 18(Suppl 1):S105–110, 2002.
- [3] Huber, W., Von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M., Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, 18(Suppl 1):S96–S104, 2002.
- [4] Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T., Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.*, 30:e15, 2002
- [5] <http://cdna01.dna.affrc.go.jp/RMOS/index.html>
- [6] <http://genome-www4.stanford.edu/MicroArray/SMD/>
- [7] NIST/SEMATECH e-Handbook of Statistical Methods, e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, 2002.
- [8] <http://www.mathworks.com/access/helpdesk/help/toolbox/stats/stats.shtml>