

# Correction Curve Normalization for Microarray Data

Tadashi Kadowaki  
PharmaDesign, Inc.  
1-1-40, Suehiro-cho, Tsurumi-ku  
Yokohama 230-0045 Japan  
kadowaki@pharmadesign.co.jp

## ABSTRACT

Microarrays have been developed for the new technology that monitors thousands of the gene expression. Recent improvement of the microarray technology is remarkable, and then the application is spread over the various regions.

However, the quality of the microarray data does not reach the former methods. Various factors of the experiment downgrade the quality of the gene expression data. For instance, the bias affects the microarray data quality. The competitive hybridization generates the bias due to the differences in the dyes (Cy5 and Cy3) incorporation. The normalization methods have been used to remove the bias, although the current method is not enough to remove all the bias.

We focus on the issue of the normalization in this paper. The usual normalization method adopts the bias as the mean value of the whole gene or housekeeping gene expression intensity. Recently, Yang *et al.*[1] and Dudoit *et al.*[2] found that usual normalization could not take some biases away. They proposed a new normalization method, where the bias depends on the spotting pin and the intensity of the spot. This normalization method is the generalization of the former methods and the accuracy is improved. In this paper, we formulate the normalization method using the correction curve. We call this method "Correction curve normalization". This normalization criterion is more general form than both the farmer and Yang's method.

We put  $X1$  and  $Y1$  are the log expression intensity of each dye measurement, and  $X0$  and  $Y0$  are the corrected log intensity. The correction curves  $f(X)$  and  $g(Y)$  convert the measurements intensity to the corrected values as,  $X0 = f(X1)$  and  $Y0 = g(Y1)$ . To rotate the coordinate  $(X, Y)$ , we have a new coordinate  $(A = (X+Y)/2, M = X - Y)$ .  $A$ -axis means the mean expression intensity and  $M$ -axis means the ratio of the sample to the control. The first-order Taylor expansion of  $M0$  around  $A1$  is expressed as,

$$M0 = f(A1) - g(A1) + \frac{1}{2}\{f'(A1) + g'(A1)\}M1 + \mathcal{O}(M1^2).$$

This equation defines the correction curve of  $M1$  to  $M0$ . Note that this correction depends on  $A1$ .

Once we estimate the correction curves, the normalization procedure is a straightforward calculation. However, the estimation of the correction curves is difficult, since the correction curve depends on the microarray slide, the sample and the control cDNA, and other experimental condi-

tions. When we assume most genes does not change between two gene expressions, the distribution of the true (or corrected) expression intensity, namely distribution of  $M0$ , should be the error distribution, the normal distribution  $N(0, \sigma^2)$ . (This assumption is reasonable for most microarray experiment that monitors a large number of genes.)

Therefore, we calculate the transformation function of the  $M1$  distribution to the  $M0$  distribution, and then we normalize the expression intensity by using this function instead of the correction curves. This normalization is based on the correction curve, so that the accuracy of the expression data should be improved and almost the same as the accuracy of the data calculated with the correction curve directly. With this normalization, the bias is removed and the variance is rescaled depending on the corresponding mean intensity  $A1$ . In general, the transformation between two distributions is non-linear, but we use linear transformation for simplicity.

The correction curve normalization is equivalent to Yang's normalization if we put  $f'(A1) + g'(A1) = 2$  for all  $A1$ . Yang's normalization removes only the bias that depends on the mean intensity  $A1$ . That is to say, the correction curve normalization includes both farmer and Yang's normalization method.

The calculation were performed by the statistical software "R"[4] and the R-package "sma" (Statistics for Microarray Analysis)[3].

## REFERENCES

- [1] Y. H. Yang, S. Dudoit, P. Luu and T. P. Speed, Normalization for cDNA Microarray Data, SPIE BiOS 2001, San Jose, California, January 2001.
- [2] S. Dudoit, Y.H. Yang, M.J. Callow and T.P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, August 2000.
- [3] <http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html>
- [4] <http://www.r-project.org>