

# A Method for Normalization of Gene Expression Data

Makoto Kano<sup>1</sup>

mkano@jp.ibm.com

Kaori Ide<sup>2</sup>

kaotide@gsc.riken.jp

Mariko Hatakeyama<sup>2</sup>

marikoh@gsc.riken.jp

Hisashi Kashima<sup>1</sup>

hkashima@jp.ibm.com

Aiko Kashihara<sup>3</sup>

kashi@spring8.or.jp

Seiki Kuramitsu<sup>3,4</sup>

kuramitu@bio.sci.osaka-u.ac.jp

Tetsuo Shibuya<sup>1</sup>

tshibuya@jp.ibm.com

Noriko Nakagawa<sup>3,4</sup>

Naka5@spring8.or.jp

Akihiko Konagaya<sup>2</sup>

konagaya@gsc.riken.jp

<sup>1</sup> Tokyo Research Laboratory, IBM Japan, 1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa 242-8502, Japan

<sup>2</sup> Riken Genomic Sciences Center, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

<sup>3</sup> Riken Harima Inst., 1-1-1 Kouto, Mikazukicho, Sayo-gun, Hyogo 679-5148, Japan

<sup>4</sup> Grad. Sch. of Sci., Osaka Univ., Machikaneyama, Toyonaka, Osaka 560-0043, Japan

**Keywords:** gene expression, normalization, probabilistic model

## 1 Introduction

Advances in microarray technologies have made it possible to comprehensively measure gene expression profiles. However, in these experiments, the efficiency of fluorescent dyes are different between different experiments, and adjustments of these expression values are mandatory before the analysis, It is referred to as normalization. Although the most popular method is total intensity normalization, where each fluorescent intensity value is divided by the sum of all the fluorescent intensities, it is difficult to apply this method in case that the expression of some of the genes changes too much. In this paper, we describe a novel method for normalizing gene expression data of microarrays so that we can compare results of microarray experiments correctly even in these difficult cases.

## 2 Method and Results

In general, expression values of majority of genes should be constant between two microarray experiments. In other words, the majority of spots at any intensity should have a ratio of 1. Our methodology is to determine adjustment factor as maximizing the expected number of genes whose expression values are constant between two experiments. Assume that  $f_{1i}$  and  $f_{2i}$  are fluorescence intensities,  $e_{1i}$  and  $e_{2i}$  are proper expression values of gene  $i$  on microarray 1 and 2 respectively.  $k$  is an offset of fluorescence intensity of microarray 2 comparing to microarray 1. Measurable fluorescence intensities are supposed to be proper expression values plus noises. Therefore, on the condition that fluorescence intensity measured is  $f_{1i}(| f_{2i})$  and an offset of fluorescence intensity of microarray 2 comparing to microarray 1 is  $k$ , the probability that proper expression value is  $e_{1i}(| e_{2i})$  is defined as follows.

$$P_1(e_{1i} | f_{1i}, k) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(e_{1i} - f_{1i})^2}{2\sigma_1^2}\right) \quad P_2(e_{2i} | f_{2i}, k) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(e_{2i} - f_{2i} + k)^2}{2\sigma_2^2}\right)$$

Note that probability distribution of noise is defined as normal distribution. Therefore, the probability of that expression value of gene  $i$  is constant between two experiments is to be

$$P(e_{1i} = e_{2i} | f_{1i}, f_{2i}, k) = \int P_1(e | f_{1i}, k) \cdot P_2(e | f_{2i}, k) de = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{\{(f_{1i} - f_{2i}) - k\}^2}{2\sigma^2}\right)$$

We determined adjustment factor as maximizing the expected number of genes whose expression values are constant between two experiments.

$$\check{k} = \operatorname{argmax} \left( \sum P(e_{1i} = e_{2i} | \hat{f}_{1i}, \hat{f}_{2i}, k) \right)$$

### 3 Results

The gene expression profiles used here are 1136 *Thermus thermophilus* genes that were measured across time-series (14 points) by cDNA microarray. Our methodology was applied for normalizing cy3 and cy5 intensities and compared with three types of conventional normalization methodologies ((A) total intensity normalization [2], (B) least square normalization and (C) lowess normalization [1]). Figure 1 is the scatter plot of fluorescent intensities of cy3 and cy5 in log space .

(1) Frequency of genes of which expression values are constant

Figure 2 is the frequency distribution of ratios and the adjustment factors determined by respective normalization methods. As this figure shows, our method could determine an adjustment factor as maximizing the number of genes whose expression values are constant, compared to conventional methods.

(2) Variance ratio

Main objective of normalization would be that (i) minimize variance of expression values that are supposed to be constant and (ii) maximize variance of expression values that are supposed to change. Therefore, desirable normalization would maximize variance ratio  $r(m)$ , which is the ratio between the average variance of top  $m$  genes and that of bottom  $m$  genes, where genes are sorted in descending order of variance values. As Table 1 shows, our method could achieve larger variance ratio than other conventional methods.

$$r(m) = \frac{\text{average of variance of top } m \text{ genes}}{\text{average of variance of bottom } m \text{ genes}}$$

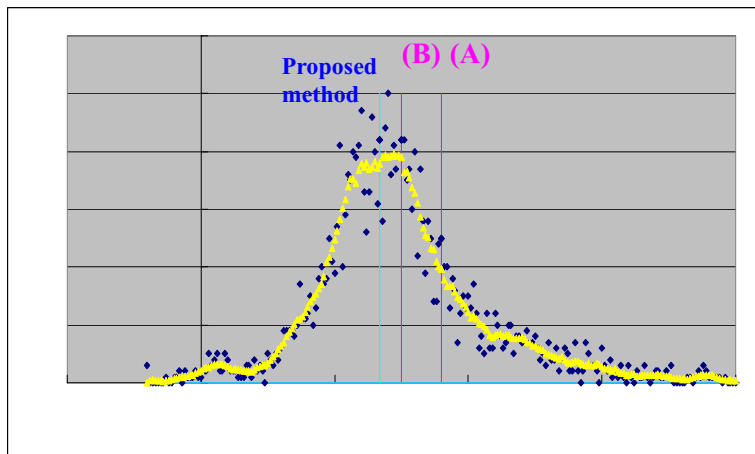
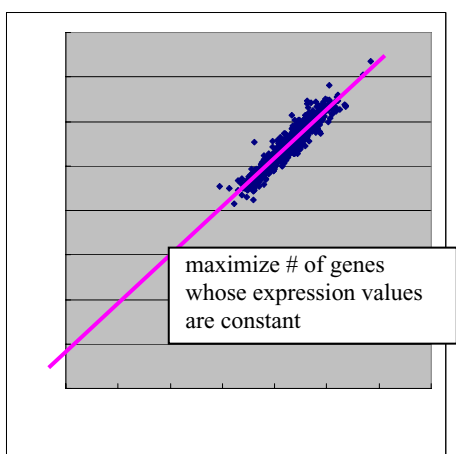


Figure 1: Scatter plot of fluorescent intensities.

Figure 2: Frequency distribution of ratios of fluorescent intensities.

Table 1: Variance ratio of each normalization method.

$m$	total intensit	least square	lowess					proposed method				
			f=0.2	f=0.4	f=0.6	f=0.8	f=1.0	$\sigma =0.02$	$\sigma =0.04$	$\sigma =0.06$	$\sigma =0.08$	$\sigma =0.1$
100	6.57	6.93	6.49	6.46	6.45	6.48	6.62	6.84	7.02	7.03	7.01	7
200	4.72	5.03	4.71	4.71	4.7	4.71	4.81	4.97	5.12	5.13	5.12	5.11
300	3.74	4.01	3.74	3.74	3.74	3.75	3.83	3.96	4.06	4.06	4.06	4.05

### References

[1] Cleveland, W.S., Robust locally weighted regression and smoothing scatterplots, *J. Amer. Stat. Assoc.*, 74:829–36, 1979.  
 [2] Quackenbush, J., Microarray data normalization and transformation, *Nat. Genet. Suppl.*, 32:496–501, 2002.