구강암 cDNA Microarray 데이터의 Normalization

김덕희[1], 최상욱[2], 박진현[*,1], 김명수[*], 이인범[1,2],
김문규[3], 김정철[3]
[*]㈜피앤아이컨설팅, [1]포항공과대학교 화학공학과,
[2]포항공과대학교 환경공학부,[3]경북대학교 의과대학교 면역학교

Normalization for cDNA Microarray Data on the oral cancer

Duk Hee Kim[1], Sang Wook Choi[2], Jin Hyun Park[*,1], Myoung Soo Kim[*], In Beum Lee[1,2],
Moon Kyu Kim[3], Jung Chul Kim[3]
[*]P&I Consulting Co., Ltd., [1]Department of Chemical Engineering, POSTECH
[2]School of Environmental Science and Engineering, POSTECH
[3]Department of Immunology, School of Medicine Kyungpook National University

INTRODUCTION

For cDNA microarray, the major sources of fluctuations can be listed according to the processes by which the data is formed.[1] Systematic variations which come into being because of the different characteristics of two dyes, background effects, overshining effects and effects resulting from image processing may be eliminated. Normalization is the general term used to represent such elimination processes.

Thousands of cDNA probes on the chip may be divided into two groups; one group is the cluster of housekeeping genes which are consistently expressed under any circumstances, and the other is that of differentially expressed genes which are discrepantly expressed depending on the condition where genes are exposed. We assume that housekeeping genes hold a large majority of the set of cDNA probes while differentially expressed genes are relatively small. For suitably normalized data, differentially expressed genes can be identified good and properly. Based on the other reported data on reproducibility, the fold difference, a criterion which discriminates differentially expressed genes from housekeeping genes, is generally not above 2 in the reliable confidence intervals.[2,3,4,5,6] Thus, reproducibility experiments were not performed, but nevertheless 2-fold difference was adopted. Goal is to modify the raw gene expression data obtained from the oral cancer cell with a suitable normalization method, to test the validation of normalization and eventually to identify differentially expressed genes that may involve in the oral cancer.

MATERIALS AND METHODS

1. cDNA microarray data and the production processes

Human 3k cDNA chip (include 3136 cDNA probes) from School of medicine, Kyungbook National University was constructed for profiling. 5’-amino-modified for attachment to glass slides, DNA clones were amplified by PCR, purified, printed automatically on the glass slides, and then hybridized with target cDNA to microarray. Two samples were exploited to obtain the data. 3136 probes were purified and reverse transcribed with fluorescently labeled cDNA. After PCR amplification, glass slide spotted with probes is prepared. Then, generated from RNAs extracted from normal (dyed with Cy3) and cancer (dyed with Cy5) cells of the oral cavity, cDNAs are hybridized. We used normal reference cells extracted from equal patient as control since a common cancerous patient has many environmental noises that we don’t want to consider in the analysis (e.g., cirrhosis and sclerosis) comparatively higher than other organs as well as other healthy persons. The microarray images were scanned by confocal laser scanner and converted in figures.

2. Intensity dependent normalization using nonparametric regression, loess method

For a spot *j*, *j* = 1, … , p, corresponding to a gene, let R_j and G_j denote the measured fluorescence intensities minus background intensities for the red (Cy5) and green (Cy3) dye, respectively. Denote *M*=ln*R/G* and *A*=ln*RG*. (ln is the natural logarithms) An *M* vs. *A* plot amounts to a 45-degree counterclockwise rotation of the (ln*G*, ln*R*)-coordinated system, multiplied by scaling factors of the coordinates. Normalization for *M* vs. *A* plot is more realistic than that for ln*G* vs. ln*R* plot because we are concerned about *R/G* ratio rather than *R* or *G* value itself. (7)

The loess (locally weighted regression scatter plot smoothing) function for multiple regression, developed by Cleveland and Devlin, was applied to the *M* vs. *A* plot obtaining for robust locally linear fit. (8) For any combination of X levels, this method fits a first-order model based on cases in the neighborhood, with more distant cases in the neighborhood receiving smaller weight. Thus, differentially expressed genes will hardly affect the loess curve. For the *i* th case, the Euclidean distance measure, denoted by d_i, is

$$d_i = \left[(X_{i1} - X_{h1})^2 + (X_{i2} - X_{h2})^2\right]^{\frac{1}{2}}. \tag{1}$$

The neighborhood about the point (X_{h1}, X_{h2}) is defined in terms of the proportion *q* of cases that are nearest to the point. d_q denotes the Euclidean distance of the furthest case in the neighborhood. The weight function used in the loess method is the tricube weight function,

$$w_i = \left[1 - (d_i / d_q)^3\right]^3 \quad d_i \le d_q \tag{2}$$

$$0 \qquad\qquad d_i \ge d_q$$

The larger is *q*, the smoother will be the fit but at the same time the greater may be the bias in the fitted value. We fixed *q* value of 0.2 that made the smallest increment of R^2 value. Given the weight for the n cases based on the distance and weight function above mentioned, weighted least squares is then used to fit a first-order linear model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \tag{3}$$

All the estimations obtained by means of the loess method were subtracted from the corresponding *M* (ln*R/G*) values.

2. 3. Normal probability plot for normality test

After normalization using the loess function, it is necessary to ascertain whether normalization gave more satisfactory results compared with results before normalization.

Each residual is plotted against its expected value under normality. Here the frequency of corrected *M* (ln*R/G*) values at equally spaced intervals corresponds to the probability density since they can be regarded as the random variable with mean 0 and estimated standard deviation $\sqrt{MSE}$. A plot that is nearly linear suggests agreement with normality, while a plot that departs fairly from linearity suggests that the residual distribution is not normal. A good approximation of the expected value of the *k*th smallest observation in a random sample of *n* is

$$\sqrt{MSE}\left[z\left(\frac{k-.375}{n+.25}\right)\right] \tag{4}$$

where $z(\cdot)$ is the inverse of the normal cumulative distribution function. (8)

As an alternative normality test, the correlation coefficient was introduced. The estimation, r_{xy} is:

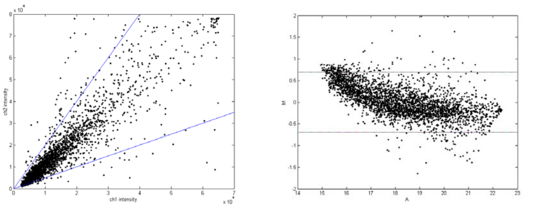
$$\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{5}$$

The closer to 1 is r_{xy}, the more linear will be the dispersion.

Result and Discussion

For housekeeping genes holding a large majority, a tendency to be linear was not shown in Figure 1. Particularly, *M* values were highly biased in the vicinity of small *A* value in *M* vs. *A* plots (Fig. 1. (b)). In fact, in spite of normalization, the distribution of *M* value is not symmetric with respect to zero. (See Fig. 3. (a)) That is, the standard deviations of the left side of the distribution is different from that of right side. This asymmetry comes from a difference in properties between two dyes, which can result from experimental variability in probe coupling and so on. (7) To solve the bias from dyes, references tried to do dye-swap experiments. In Fig. 1. (c), hundreds of genes considered to be definitely housekeeping exceeded the line of 2-fold difference, which indicated that incorrect genes may be selected without normalization. For the large G and R values, some genes were concentrated, which is inferred as an error from image scanning. The loess curves for the entire data sets are shown in Figure 2.

In raw data without normalization, the number of genes having *M* value larger than ln(2) in data set #1 and #2 were 148 and 28, respectively, and subsequently 14 common genes were found. (See Table 1.) On the other hand, the number of genes having values above two fold difference in normalized data was almost the same. i.e., 47 on data set #1 and 45 on #2. As mentioned above, on data set #1 before normalization, on the ground of biased *M* values, many of 148 genes on data set #1 were incorrectly selected without normalization, whereas they were eliminated after normalization. We may have confidence that genes having *M* value larger than ln(3) and smaller than ln(1/3) are confidentially differentially expressed genes if they included no fluctuations. As a result, the genes having biological meanings were identified.



(a) Data set #1

(b) *M* vs. *A* plot - Data set #1

Figure 1. Scatter plots without normalization. Ch1 indicates the reference intensities, *G*, and Ch2 the control, *R*. (a) is raw data plots, and (b) is *M* vs. *A* plots. Lines indicate two-fold difference.

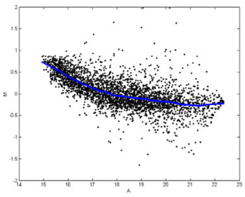
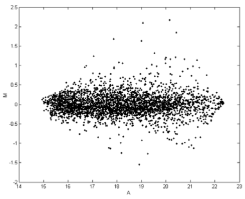
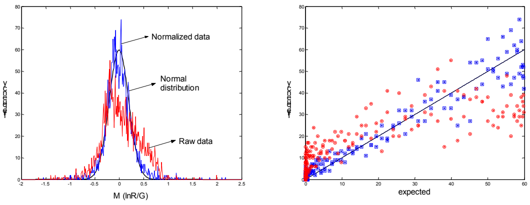


Figure 2. *M* vs. *A* plots for normalization with the loess lines (the blue in (a)). (b) is the plot which the loess lines were set for zero.

(a) Data set #1 and its loess curve (left)

(b) Normalized Data set #1 (right)





(a) M vs. frequency plot (b) q-q plot

Figure 3. Normal probability plot of data set #1. In (a), red line indicates the frequency of genes having corresponding *M* values in raw data, blue line in normalized data. (b) is q-q plot of lines in (a). When the points are close to the line y=x, the distribution is said to be normal. At a glance, it can be realized that normalized data (the blue points) are located in nearby y=x, while raw data (the red) are skewed left.

Rawdata (*M*)	0329	1228	Common genes
< ln(1/2)	64	122	30
> ln(2)	148	28	14

(a) in raw data

Normalized data (*M*)	0329	1228	Common genes
< ln(2/3)	150	308	74
< ln(1/2)	35	58	17
< ln(1/3)	6	11	2
> ln(3/2)	174	216	67
> ln(2)	47	45	17
> ln(3)	13	11	6

(b) in normalized data

Table 1. The number of candidates for differentially expressed genes in raw and normalized data.

REFERENCES

1. Normalization strategies for cDNA microarrays
2. Determining significant fold differences in gene expression analysis
3. Anonymous. GEM Microarray Reproducibility Study. Incyte Pharmaceuticals, Inc. (1999).
4. F. Bertucci, et al., “Expression scanning of an array of growth control genes in human tumor cell lines” Oncogene. 18:3905-12 (1999).
5. G. K. Geiss, et al., “Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays” Virology. 266:8-16 (2000).
6. C. K. Lee, R. Weindruch, and T. A. Prolla, “Gene-expression profile of the ageing brain in mice” Nat Genet. 25:294-7 (2000).
7. Normalization for cDNA Microarray Data
8. Applied linear statistical models fourth editio