Development of a Reproducibility Index for cDNA Microarray Experiments

Byung Soo Kim[1], Sun Young Rha[2]

Abstract

Since its introduction in 1995 by Schena et al. cDNA microarrays have been established as a potential tool for high-throughput analysis which allows the global monitoring of expression levels for thousands of genes simultaneously. One of the characteristics of the cDNA microarray data is that there is inherent noise even after the removal of systematic effects in the experiment. Therefore, replication is crucial to the microarray experiment. The assessment of reproducibility among replicates, however, has drawn little attention. Reproducibility may be assessed with several different endpoints along the process of data reduction of the microarray data. We define the reproducibility to be the degree with which replicate arrays duplicate each other. The aim of this note is to develop a novel measure of reproducibility among replicates in the cDNA microarray experiment based on the unprocessed data. Suppose we have p genes and n replicates in a microarray experiment. We first develop a measure of reproducibility between two replicates and generalize this concept for a measure of reproducibility of one replicate against the remaining n-1 replicates. We used the rank of the outcome variable and employed the concept of a measure of tracking in the blood pressure literature. We applied the reproducibility measure to two sets of microarray experiments in which one experiment was performed in a more homogeneous environment, resulting in validation of this novel method. The operational interpretation of this measure is clearer than Pearson's correlation coefficient which might be used as a crude measure of reproducibility of two replicates.

Key words: cDNA microarray; Reproducibility; Duplication.

1. Introduction

Since its introduction in 1995 by Schena et al. (1995) DNA microarrays have been established as a tool for high-throughput analysis which allowed the global monitoring of expression levels for thousands of genes simultaneously. As Gibson (2002) noted, they are much more than a tool when used in conjunction with classical genetics and the emerging

1) Professor, Dept. of Applied Statistics, Yonsei University, Seoul, 120-749.
2) Assistant Professor, Cancer Metastasis Research Center, College of Medicine, Yonsei Univerisity, Seoul, 120-749

new field of bioinformatics, bcause they can force us to change our perspective on the process under study. There are two widely used technologies, cDNA microarrays and oligonucleotide arrays depending on how nucleic acid probes are arrayed at high density for interrogation of labeled mRNA samples. We focus on the cDNA microarray in this note. However, the same procedure will apply in both technologies.

The cDNA microarray experiment is initiated by spotting and immobilizing DNA on a glass slide or on other substrate, the microarrays in batches of perhaps 100 slides. Next, mRNA isolated from the cells or tissues under study is reverse-transcribed into a cDNA and one of two fluorescent dye labels, cy3 and cy5, is incorporated. The dye-labelled cDNA can hybridize with complementary sequences on the array. Then the array is scanned for cy3 and cy5 fluorescent intensity. This competitive hybridization of two samples labeled with these two dyes allows an estimate of the ratio of transcript abundance in terms of fluorescent intensity. Further, if we assume that the hybridization efficiency of an individual DNA clone is not affected by the dye level, the relative abundance of a mRNA in the two sample can be measured by the relative intensity of two fluorescents. The idea is that for most genes, steady state mRNA levels approximate protein levels and thus quantitative expression monitoring at the mRNA level provides important clues as to function (Schena and Davis, 1999).

One of the characterizing properties of the cDNA microarray data is that there is inherent noise even after the removal of systematic effects in the experiment. Therefore, replication is crucial to microarray experiment (Jin et al., 2002; Lee et al., 2000; Kerr et al, 2001; Tseng et al., 2001). As noted by Kerr and Churchill (2001) there are several levels of replications: genes can be spotted multiples times per array, mRNA samples can be used on multiple arrays, and mRNA samples can be taken from multiple specimens to account for inherent biological variability. Among these three levels we slightly modify the second concept and refer in this note to replicates as multiple arrays with the same pool of total RNA. Lee et al. (2000) elegantly described the importance of replication in the microarray studies and confirmed that replication was not equivalent to duplication and hence was not a waste of scientific resources. However, the assessment of reproducibility among replicates, i.e., how much replication is close to duplication, has drawn little attention.

Reproducibility among replicates may be assessed with several different endpoints along the process of data reduction of the microarray data analysis. Yue et al. (2001) described various aspects of the experimental parameters, including the amount of target DNA spotted on the glass slide and the amount of mRNA as input for labeling reactions, for the manufacture of highly reproducible cDNA microarrays. They first reported the optimal values of parameters and demonstrated the reproducibility of the optimized microarray experiment by observing that a summary measure, e.g., the coefficient of variation of the differential expression, showed a stable value across different tissue types and array batches. Kerr, Martin and Churchill (2000) employed two different designs for the microarray experiment and performed analysis of variance to obtain estimates and their confidence intervals of interaction terms (VG) under each design. They then constructed a 3x3 contingency table based on the numbers of up-regulated, down-regulated and unaffected genes under each design and computed the concordance rate to demonstrate the reproducibility of estimated changes in expression levels under two designs.

The aim of this paper is to develop a measure of reproducibility among replicates in the cDNA microarray experiment based on the unprocessed data.

2. Methods and Results

Suppose we have p genes and n replicates in a microarray experiment. Let Y_{ij} denote the logarithm of intensity of two dyes for i = 1, ···, p and j = 1, ···, n. Reproducibility is defined to be a degree with which two replicate arrays duplicate each other. Perfect reproducibility is the duplication and no reproducibility means that two arrays are statistically independent. We first develop a measure of reproducibility between two replicates. Then we generalized this concept and propose a measure of reproducibility of one replicate against the remaining n-1 replicates.

Let R_{ij} be a rank assigned to Y_{ij} among $\{Y_{1j}, \cdots, Y_{pj}\}$. Define $r_{ij} = R_{ij}/p$ and $D_{jj'}(i) = r_{ij} - r_{ij'}$ for $i = 1, \ldots, p$ and $j, j' = 1, \ldots, n$. For a given δ ($0 \le \delta \le 1$) define for $j \neq j'$;

$$f_{jj'}(\delta) = \frac{1}{p} \sum_{i=1}^{p} I_{(-\delta, \delta)}(D_{jj'}(i)), \qquad (1)$$

where $I_A(x)$ is an indicator function such that $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$. $f_{jj'}(\delta)$ is the proportion of genes in arrays j and j′ for which the difference between percentiles of Y_{ij} and $Y_{ij'}$ is smaller than δ. $f_{jj'}(\delta)$ is an increasing function of δ and $f_{jj'}(\delta)$ converges to 1 as δ tends to 1.

Now, define $F_{jj'}$ as in equation (2).

$$F_{jj'} = \int_0^1 f_{ij'}(\delta) d\delta \qquad (2)$$

Under the assumption of no reproducibility between arrays j and j′ we may assume that r_{ij} and $r_{ij'}$ are independent and identically distributed (i.i.d) random variables with uniform distribution over the interval [0,1]. Therefore, even under no reproducibility between two replicates there is positive probability $f_2(\delta)$ defined in equation (3) that the event $\{|r_{ij} - r_{ij'}| < \delta\}$ occurs.

$$f_2(\delta) = \Pr[\{|r_{ij} - r_{ij'}| < \delta\}] \qquad (3)$$
$$= \delta(2 - \delta)$$

Define F_2 as the integral of $f_2(\delta)$ over [0,1] as in equation (4).

$$F_0 = \int_0^1 f_0(\delta)\, d\delta = \int_0^1 \delta(2 - \delta)\, d\delta = \frac{2}{3} \qquad (4)$$

We propose the index.R in equation (5) for a measure of reproducibility between two replicates.

$$\text{Index.R} = \frac{F_{jj'} - F_0}{1 - F_0} \qquad (5)$$
$$= 3\left(F_{jj'} - \frac{2}{3}\right)$$

Index.R tells us how close the reproducible feature, if there exists, of two replicates is towards the duplication after the reproducibility by chance alone is adjusted.

When we would like to measure the reproducibility of a replicate array j against the remaining (n-1) replicates, we may envisage an imaginary array j′ of which its percentile $r_{ij'}$ is represented by

$$r_{ij'} = \frac{1}{n-1} \sum_{s \neq j} r_{is} \qquad (6)$$

We have the same equations (1) and (2) for defining $f_{jj'}$ and $F_{jj'}$, respectively. But, we need to modify the probability that $\{| r_{ij} - r_{ij'}| < \delta\}$ occurs under no reproducibility condition, because $r_{ij'}$ is now the mean of (n-1) i.i.d uniform [0,1] random variables.

Following the result in Johnson and Kotz (1970, equation 19 of page 64) it can be shown that

$$F_n = \int_0^1 \Pr[\{| r_{ij} - r_{ij'}| < \delta\}]\, d\delta \qquad (7)$$

$$= \begin{cases} \frac{2}{3} & n = 2 \\ \frac{17}{24} & n = 3 \\ \frac{13}{18} & n = 4 \\ \frac{35}{48} & n = 5 \end{cases}$$

where $r_{ij'}$ is defined in equation (6)

Now, we generalize index.R of equation(5) for the case of n replicates as

$$\text{index.R} = \frac{F_{ij'} - F_n}{1 - F_n} \qquad (8)$$

We refer to index.R's of equations (5) and (6) as pairwise index.R and overall index.R, respectively.

In the presentation we validate the index.R as a measure of the reproducibility using two cDNA microarray experiments in which one was performed in a more homogeneous environment.

References

Gibson G. (2002). Microarray in ecology and evolution: a preview. *Molecular Ecology* 11:17-24.

Jin W, Riley RM, Wolfinger RD, Whilte KP, Passador-Gurgel G, Gibson G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster. Nature Genetics* 29:389-395.

Johnson NL, Kotz S. (1970). *Distributions in statistics, continuous univariate distributions-2*, Wiley: New York

Kerr MK, Afshari CA, Bennett L, Bushel P, Martinez J, Walker N, Churchill GA. (2001). Statistical analysis of a gene expression microarray experiment with replication.

To appear in *Statistica Sinica*, also www.jax.org/research/churchill/pubs/index.html.

Kerr MK, Churchill GA (2001). Statistical design and the analysis of gene expression microarrays, *Genetical Research* 77:123-128.

Kerr MK, Martin M, Churchill GA (2000). Analysis of variance for gene expression microarray data, *Journal of Computational Biology* 7:819-837.

Lee M-LT, Kuo FC, Whitmore GA, Sklar J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive DNA hybridization. *Proc. Natl. Acad. Sci.* 97:9834-9839.

Schena M, Davis RW, (1999). Genes, genomes and chips. In Schena M.(ed). *DNA microarrays*, Oxford University Press: Oxford, pp.1-15.

Schena M, Shalon D, Davis RW, Brown PO. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* 270:368-9.

Tseng GC, Oh M-K, Rohlin L, Liao JC, Wong WH. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research* 29:2549-2557.

Yue H, Eastman PS, Wang BB, Minor J, Doctorlero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R. (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Research* 28(8):e41 (9 pages).