Bioinformatics and Biosystems 2006, Vol. 1, No. 1, pp. 67-72

Effect of missing values in detecting differentially expressed genes in a cDNA microarray experiment

Byung Soo Kim[1], Sun Young Rha[2,3]

[1]Dept of Applied Statistics, Yonsei University, Seoul, 120-749, Korea. [2]Cancer Metastasis Research Center, College of Medicine, Yonsei University, Seoul, 120-752, Korea. [3]Brain Korea 21 Project for Medical Science, College of Medicine, Yonsei University, Seoul, 120-752, Korea.

Abstract

The aim of this paper is to discuss the effect of missing values in detecting differentially expressed genes in a cDNA microarray experiment in the context of a one sample problem. We conducted a cDNA microarray experiment to detect differentially expressed genes for the metastasis of colorectal cancer based on twenty patients who underwent liver resection due to liver metastasis from colorectal cancer. Total RNAs from metastatic liver tumor and adjacent normal liver tissue from a single patient were labeled with cy5 and cy3, respectively, and competitively hybridized to a cDNA microarray with 7775 human genes. We used $M=\log_2(R/G)$ for the signal evaluation, where R and G denoted the fluorescent intensities of Cy5 and Cy3 dyes, respectively. The statistical problem comprises a one sample test of testing $E(M)=0$ for each gene and involves multiple tests. The twenty cDNA microarray data would comprise a matrix of dimension 7775 by 20, if there were no missing values. However, missing values occur for various reasons. For each gene, the no missing proportion (NMP) was defined to be the proportion of non-missing values out of twenty. In detecting differentially expressed (DE) genes, we used the genes whose NMP is greater than or equal to 0.4 and then sequentially increased NMP by 0.1 for investigating its effect on the detection of DE genes. For each fixed NMP, we imputed the missing values with K-nearest neighbor method (K=10) and applied the nonparametric t-test of Dudoit et al. (2002), SAM by Tusher et al. (2001) and empirical Bayes procedure by Lönnstedt and Speed (2002) to find out the effect of missing values in the final outcome. These three procedures yielded substantially agreeable result in detecting DE genes. Of these three procedures we used SAM for exploring the acceptable NMP level. The result showed that the optimum no missing proportion (NMP) found in this data set turned out to be 80%. It is more desirable to find the optimum level of NMP for each data set by applying the method described in this note, when the plot of (NMP, Number of overlapping genes) shows a turning point.

Introduction

The DNA microarray has been established as a major tool in biological researches due to its ability of monitoring gene expression levels of thousands of genes simultaneously under different conditions (Jin et al., 2001: Gibson 2002; Hedenfalk, 2002; Olesiak 2002; Ramaswamy 2002; Huang 2003; Keshave and Ong, 2003). It is not trivial to analyze the data from microarray experiment, not because they just involve large amount of data, but because they comprise a non-standard statistical problem which is often referred to as a "large p, small n" problem (West, 2003). Typically, we have thousands of genes (=p) for a microarray experiment with tens of microarrays (=n). Several analysis tools including SAM (Tusher et al, 2001) and BRB-ArrayTools (Simon and Peng) have been introduced in the pubic domain to provide a guidance to laboratory scientists on the statistical analysis of microarray data.

Microarray experiment data can be represented by a p x n matrix, where the (i, j)th element of the matrix indicates the i-th gene expression level for the j-th microarray, i=1, .., p, and j=1, ···,n. It is quite often that we observe the missing values in the data of p x n matrix. Missing values occur for various reasons not only from the technical problems but also

Corresponding author: Byung Soo Kim (Tel: +82-2-2123-4541, Fax:+82-2-313-5331, Email: bskim@yonsei.ac.kr)
B.S. Kim's study was supported by Yonsei University Research Fund of 2001. S.Y. Rha's study was supported by a grant of the IMT-2000 project, Ministry of Health & Welfare, Republic of Korea (01-PJ11-PG9-01BT00A-0028).

from the biological characteristics. Currently, as the chip quality and the hybridization techniques have been improved to a certain level, the missing values usually come from the biological reasons, such as no expression of the specific genes in the sample or the inefficient dye labeling.

These missing data are usually excluded from the subsequent analysis. Sometimes, missing values are replaced by 0's after logarithmic transforming of the data or by a row average, when the researcher finds it difficult to conduct a replicate experiment as a clear solution to the problem. As is indicated in Troyanskaya et al. (2001), this approach is not optimal, since these methods do not utilize the correlation structure of the data. Troyanskaya et al. (2001) recommended the K-nearest neighbor (KNN) method for imputing missing values in microarray data as a result of their comparative study of three methods including the singular value decomposition (SVD) based method and row average. Following Troyanskaya et al's recommendation, we use KNN method for the imputation of missing values.

Before we discuss the imputation of missing values we need to determine, however, whether a certain gene which has m missing values in a set of n arrays should be included in the analysis. We define no missing proportion (NMP) of a gene which has m missing values out of n arrays as (n-m)/n. While we analyzed a set of microarray data, we found that a set of differentially expressed (DE) genes varied substantially depending on the NMP. Alizadeh et al. (2000) used 80% NMP for a microarray experiment consisting of 96 arrays. However, they didn't provide any justification of the 80% figure.

In this study we investigate the effect of missing values on the detection of DE genes and suggest a method to find an optimum value of NMP.

Materials and Methods

Microarray experiment and Data

We performed cDNA microarray experiments based on 20 patients who underwent liver resection due to liver metastasis from colorectal cancer from September to December 2001 at Severance Hospital, Yonsei Cancer Center, Yonsei University College of Medicine, Korea. The cDNA microarray experiment was done with microarrays spotted with 7775 sequence-verified human cDNA(Genomic Tree, Korea) following the in-house protocol of Cancer Metastasis Research Center, Yonsei University College of Medicine, Korea (in submission). Briefly, 50㎍ of total RNAs' from metastatic tumor and adjacent normal liver tissues from a single patient were labeled with cy5 and cy3 by reverse transcription using the Superscript II enzyme (Invitrogen, U.S.A), respectively. cDNA microarray was prehybridized in 3.5X SSC, 0.1% SDS, 10mg/ml BSA for 1 h at 42℃ before hybridization with labeled targets. Cy5- and Cy3- labeled targets were mixed with 30μg human Cot-1 DNA, 20μg poly (dA)-poly (dT), and 100μg yeast tRNA. A Microcon-30 filter (Amicon, MA, USA) was used to purify and concentrate the hybridization mixture, which was then adjusted to 3.4X SSC and 0.3% SDS in a final volume of 90㎕. Following denaturation at 100℃ for 1.5 minutes and 30 minutes of pre-annealing at 37℃, the labeled target was hybridized to the array at 65℃ for 16-24 hours. The slide was then washed for 2 min each in 0.5X SSC/0.01% SDS, 0.06X SSC/0.01% SDS, and 0.05X SSC, consecutively, at room temperature and spun-dried before scanning. Hybridized arrays were scanned using a GenePix 4000B (Axon Instruments, USA).

$M=\log_2(R/G)$ is used for the signal evaluation. For each gene, say for the g-th gene, we would like to test the following hypothesis based on 20 observations

$$H_0(g) : E(M)=0 \quad \text{vs} \quad H_a(g) : E(M)\neq 0. \qquad (1)$$

This constitutes a one-sample problem with multiple tests. Data can be represented by a 7775 x 20 matrix when there were no missing values. However, missing values occur for various reasons. Genes with low NMP would be excluded from the analysis. We counted the number of genes whose NMP is greater than or equal to x, where x ranges from 0.4 up to 1.0 with an increment of 0.1, displayed in Table 1. It may be noted that the result in Table 1, which showed a decreasing trend of the proportion of valid genes as NMP went up, represented the characteristic of NMP in general, not necessarily a characteristic of the dataset at hand.

Table 1. The number of genes whose no missing proportion(NMP) is greater than or equal to x (x=0.4~1.0)..

No missing proportion	0.4	0.5	0.6	0.7	0.8	0.9	1
Number of genes	5643	5043	4335	3627	3012	2376	1472
% out of total	72.6	64.9	55.8	46.6	38.7	30.6	18.9

Statistical Procedures

We apply following three procedures for the one-sample problem of Equation (1) to detect a set of DE genes for each NMP.

1) Dudoit et al.'s nonparametric t test with Westfall and Young's step-down method for the p-value adjustment

Dudoit et al. (2002) employed the family-wise error rate (FWER) for controlling the type I error and used Westfall and Young's step down procedure for calculating the adjusted p-value. We have 2^{20} permutations to derive the nonparametric null distribution of a one sample t statistic. We developed a C++ program for the derivation of the adjusted p-value of the t-statistic and investigated the number of DE genes for several different values of FWER (0.01, 0.05 and 0.001).

2) Tusher et al.'s SAM procedure

Tusher et al's SAM procedure is a permutation test with a modified t statistic. They adopted the false discovery rate (FDR, Benjamini and Hochberg, 1995) for controlling the type I error, where FDR is defined to be the number of false positive genes divided by the number of declared significant genes. FDR is more sensitive in detecting significant genes (Ge et al. 2003). We used Tusher et al's SAM software program to detect the DE genes. We set K=10 when we use KNN method for imputing missing values and fix 5000 for the number of permutation in running SAM program. Delta value in SAM was determined by taking it into consideration that our lab can take about 100 DE genes for the further confirmatory experiment like RT-PCR.

3) Lönnstedt and Speed's empirical Bayes procedure

Lönnstedt and Speed (2002) used empirical Bayes method to derive a Bayes log posterior odds, say B statistic. One may take top 50, 100 or 150 genes in terms of B values in combination with experimental preference.

Results

The numbers of DE genes detected by Dudoit et al.'s nonparametric t test corresponding to FWER=0.01, 0.005 and 0.001 for NMP=1.0 are given in Table 2. Lönnstedt and Speed's B statistic provided a similar list of top 100 DE genes with Dudoit et al.'s result. We could detect 100 DE genes by adjusting delta value in SAM procedure. These three procedures yielded substantially agreeable results in detecting DE genes as one can find in Figure 1.

Table 2. The number of differentially expressed (DE) genes detected by Dudoit et al's t-test for FWER=0.01, 0.005 and 0.001 and no missing proportion (NMP)=1.0.

FWER	0.01	0.005	0.001
# of significant genes	411	344	160
total # of genes	1472 (NMP=1.0)		

As we can adjust delta value to find a pre-fixed number of DE genes, we used SAM for identifying the optimum NMP. We could detect 75 DE genes for NMP=1.0 using SAM procedure and the q-value (the smallest FDR at which a gene is called significant) for each gene was almost 0. We further applied SAM for detecting approximately 75 DE genes by decreasing the NMP by 0.1 down to 0.4. For each NMP the number of overlapping genes with the set of DE genes for NMP=1.0 is listed in Table 3.

Table 3. The number of overlapping genes under no missing proportion (NMP) less than 1.0 with the set of differentially expressed (DE) genes corresponding to NMP=1. +(-) means positively (negatively) significant genes.

NMP(%)	Overlapping genes		
	+	-	total
100	28	47	75
90	23	31	54
80	21	23	44
70	23	23	46
60	20	18	38
50	21	13	34
40	20	13	33

The number of overlapping genes as a function of no missing proportion from Table 3 is exhibited in Figure 2, from which one can detect a turning point of the slope at NMP=0.8. We propose this value is the optimum NMP for the microarray analysis.

Discussion

As the knowledge on biology is improving, it is proven that the complex integrative interactions of many molecules are essential in most of biological process. Hence the high-throughput technology such as microarray giving thousands of gene expressions simultaneously under the specific condition becomes one of the most useful tools to understand the pathophysiology of the various disease

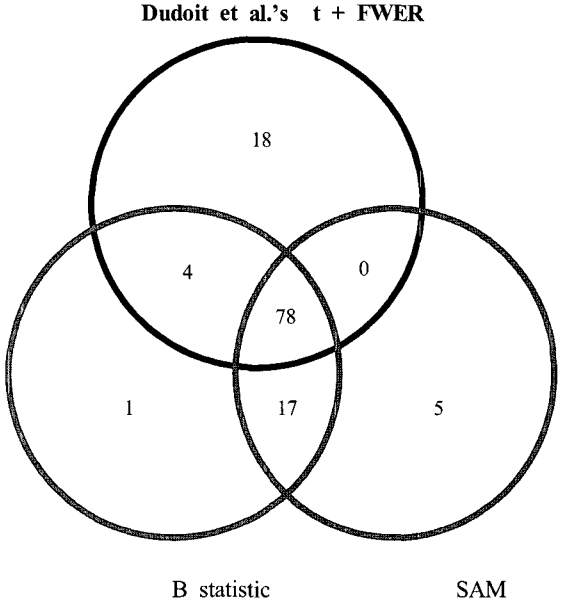

Figure 1. The overlap pattern of top 100 differentially expressed genes detected by three procedures; t test of Dudoit et al., Tusher et al's SAM, and Lönnstedt and Speed's B statistic. The number in the intersection indicates the number of genes jointly detected by two or three procedures. The FDR of SAM is 0.0017. The log posterior odds ratio of the top 100^{th} gene in the Lönnstedt and Speed's B statistic is 6.29 and the adjusted p-value of the top 100th gene of Dudoit et al.'s t test with FWER approach is less than 0.001.

conditions including cancer. Because cancer is fatal when the disease is metastasized to other organ from its original tissue, we designed this study to identify the genomic characterization of colorectal cancer with liver metastasis. Our experiment was conducted in direct design since we hybridized a metastatic tumor and a normal tissue competitively in a single array. Usually a dye-swap experiment of the same number of arrays needs to be performed to balance the dye effect and hence minimize the possible bias in comparison of tumor and normal tissue. We didn't perform this dye-swap experiment, which might give bias in the detection of DE genes. However, we don't believe it may have any effect on the general conclusion of this note, considering the improved hybridization techniques and the reproducible quality of microarray data used in the analysis.

There are many issues to solve in using microarray from technical, biological, and analytical point of view. As the biotechnology is improving, the technical problems of qualified chips and hybridization techniques have been solved. Meanwhile, somewhat interactive biological and analytical

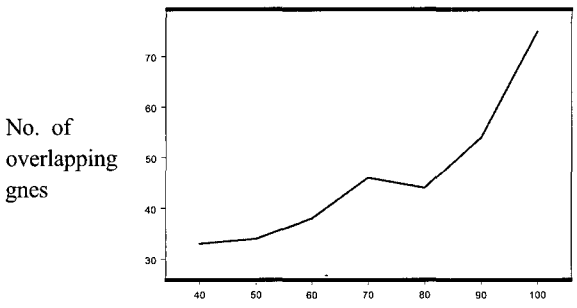

Figure 2. The number of overlapping genes as a function of no missing proportion. One can detect a turning point of the slope at no missing proportion (NMP)=0.8, which we propose to be the optimal NMP for the microarray analysis.

problems still cause many troubles, including experimental design, requirement of replicates, selecting the differentially expressed genes, method of pattern recognition and the validation of selected genes. Fortunately, the potential suggestions of these problems are introduced and under the evaluation in many ways. However, there is no considerable concern on how to handle the missing values, which is one of the basic issues in microarray data analysis. Recently, as the hybridization technique is improved, the missing values usually develop from the biological reasons, such as no expression of the specific genes in the sample or inefficient dye labeling. These missing data are usually excluded from the subsequent analysis. However, the use of rare clinical samples with various RNA expression profiling, requires more improved analysis to handle the missing data adequately without losing many data points.

In selecting DE genes, Callow et al. (2000) reported only 8 significant genes out of 5548 genes which had altered expression in a mouse model (Apo A1 k.o) with very low HDL cholesterol levels compared to inbred control mice. In contrast to Dudoit et al's report, we believe this number of significant genes is quite plausible even with the conservative FWER controlled test, when we consider the heterogeneity of tumors and the physiological difference between the normal tissue and tumor. When we compared the gene list of selected genes by 3 procedures of Dudoit et al's, Lönnstedt and Speed's B statistic's, and SAM, three procedures yielded substantially agreeable results in detecting DE genes (Figure 1).

In Table 3 we note the number of positive genes shows a stable trend for NMP $\leq$ 0.9, whereas the number of negative genes exhibits a decreasing trend as NMP goes down. This may pose a biological question which may warrant further investigation.

We found that SAM was particularly useful for our purpose of investigating the effect of missing proportions on the detection of DE genes, because by adjusting the delta value one could find a pre-fixed number of DE genes under different data sets. Thus, we restrict ourselves to SAM for the further analysis. The optimum NMP found in this metastatic colorectal cancer data set turned out to be 80%, which happened to coincide with Alizadeh et al's NMP. It may be hard to generalize that this 70~80% NMP would be the acceptable NMP level. It is more desirable to find the optimum level of NMP for each data set by applying the method described in this note. When the plot of "NMP-Number of overlapping genes" does not show a turning point, the researcher should develop another method to determine the optimal level of NMP.

References

[1] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A,Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Scherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Bryd JC, Botstein D, Brown PO, Staudt LM. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403:503-511.

[2] Callow MJ, Dudoit S, Gong EL, Speed TR, Rubin EM (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res. 10:2022-2029

[3] Dudoit S, Yang YH, Callow MJ, Speed TP. (2002). Statistical methods for differentially expressed genes in replicated cDNA microarray experiments. *Statistica* Sinica 12:??

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

[5] Keshava N and Ong T. (2003). Gene expression patterns in human liver cells exposed to tetrachloroethylene and its metabolite using microarray analysis. *Environ. Mol. Mutagen.* 41:182.

[6] Hedelfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Metzler P, Gusterson B, Esteller M, Kallioniemi O-P, Wifond B, Borg A, Trent J. (2001). Gene-expression profiles in hereditary breast cancer, New. *Engl. J. Med.* 344:539-548.

[7] Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'AMico M, Pestell RG, West M, Nevin JR. (2003). Gene expression phenotypic models that predict oncogenic pathway. *Nat. Gen.* 34:226-230.

[8] Jin L, Riley RM, Wolfinger RD, White KP, Pasador-Gurgel G, Gibson G. (2001).The contribution of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Gen. 29:389-395.

[9] Olesiak MF, Churchill GA, Crawford DL. (2002). Variation in gene expression within and among natural populations. Nat. Gen. 32:261-266.

[10] Lönnstedt I, Speed T. (2002). Replicated microarray data. *Statistica Sinica* 12:31-46.

[11] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang C-H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Geral W, Loda M, Lander ES, Golub TR. (2001) Multiclass cancer diagnosis using tumor gene expression signature. *Proc. Natl. Acad. Sci.*, 98:15149-15154.

[12] Simon R, Peng A. (2002). BRB-ArrayTools Version 2.1. Biometric Research Branch, NCI, U.S.A.

[13] Simon R, Radmacher MD, Dobbin K. (2002). Design of studies using DNA microarrays. Genetic *Epidemiology* 23:21-36.

[14] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520-525.

[15] Tusher V, Tibshirani R, Chu G. (2000). Significance analysis of microarray applied to the ionizing radiation response. *Proc. Nat. Acad. Sci.*, 98:5116-5121.

[16] West M. (2003). Bayesian factor regression models in the "Large p, Small n" paradigm. in: J.M. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West (eds) Bayesian Statistics 7, Oxford University Press, pp. 723-732.