Detecting differentially expressed genes from a mixed data set

Sunho Lee[1], Inyoung Kim[2], Sangcheol Kim[2],
Sun Young Rha[2], Hyun Chel Chung[2], Byung Soo Kim[3]

Abstract

When we have both a paired data set and two independent data sets, neither a paired t-test nor a two-sample t-test can be used to detect differences between two samples. In order to identify differentially expressed genes in a mixed data set, a new test statistic is proposed.

KEY WORDS : microarray experiment, gene expression, a mixed data set, t-test, discriminant analysis

1. Introduction

This research is a part of the on-going project in which we identify a set of differentially expressed (DE) genes in colorectal cancer, compared with normal colorectal tissues, to evaluate its predictivity of a new specimen and eventually to rank genes for the development of biomarkers for population screening of colorectal cancer.

In case of using clinical samples, fresh collection of the tissue samples and the adequate storage are essential, especially for RNA preparation. In addition, regardless of the proper banking process, some type of tissues are very fragile for RNA, such as pancreas tissue or gastrointestinal tract tissue which are related to enzymatic activity in tissue itself, resulting in the degradation of tissue RNA spontaneously. In case when the tumor area or the adjacent normal tissue is small, it is hard to banking the tissues. With all these possible causes, even that we designed the study for paired tissue samples, it happens occasionally not to be able to prepare all the samples in pairs as we planned, resulting in 'a mixed data set' of paired samples and independent normal or

[1] Department of Applied Mathematics, Sejong University, Seoul, Korea.
This work was supported by grant R04-203-000-10145-0 from the Basic Research Program of the Korea Science and Engineering Foundation.
[2] Cancer Metastasis Research Center, Yonsei University College of Medicine, Seoul, Korea.
This work was supported by the Korea Science and Engineering Fund through the Cancer Metastasis Research Center at Yonsei University.
[3] Department of Applied Statistics, Yonsei University, Seoul, Korea.
This work was supported by a grant of the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (02-PJ1-PG3-10411-00-03).

tumor tissues. As the clinical samples are very precious, especially with long-term clinical follow-ups, and the limited numbers of clinical samples, it is necessary to find out the way to use all the possible data set in analysis.

In statistical analysis for comparing two different samples, either a paired t-test for paired samples or a two-sample t-test for two independent samples are frequently used. But, when we have a mixed data set, there is no standard test to deal with. Therefore, sometimes, paired samples in a mixed data set are treated as two independent samples[1,2] to apply a two samples t-test, or two independent samples are dropped, if the cases are small, to apply a paired t-test. Both cases are failed to utilize all the information that the data set has and necessity to construct a test statistic to use the whole data set is arisen.

With the basic idea of combining a paired t-test with a two samples t-test, we establish a new test statistic, t_3, which is applicable to three data types of matched pairs and two independent data sets. We used the high density microarray data to identify the significant classifier genes to differentiate cancer from normal tissues in colorectal cancer patients, which have both paired and unpaired data.

2. A mixed data set

We collected cancers and normal tissues from 87 colorectal cancer patients during the operation and snap froze the tissues at –70C. We attempted to extract total RNAs from both tumor and normal tissues for each patient, but we failed to extract RNAs from tumor tissues for 19 patients and from normal tissues for 32 patients, respectively. We conducted a cDNA microarray experiment using a common reference design with 17K human cDNA microarrays.

Finally, from each of 36 patients we had RNA samples both for tumor and normal tissues. From 19(32) patients RNA samples for normal (tumor) tissues only were available. Thus, a mixed data set with a matched pair sample of size 36 and two independent samples of sizes 32 and 19 are produced. After simply removing the flagged spots, we also removed the genes that the signals are missing in more than 20% of the samples. Then, the missing values were adjusted with k-NN method. Within-print tip group intensity dependent method [3] is adopted for the normalization of log intensity ratio.

3. Detecting differentially expressed genes

Let X_{ij} and Y_{ij} be the log expression levels for the i^{th} gene of the j^{th} patient's tumor and normal samples, respectively. For the first n_1 patients, we have paired log expression levels for

both tumor and normal samples, (X_{ij}, Y_{ij}), $j = 1, \cdots, n_1$. For the next patients, we get n_2 log expression levels of tumor only samples, X_{ij}, $j = n_1 + 1, \cdots, n_1 + n_2$ and n_3 log expression levels of normal only samples, Y_{ij}, $j = n_1 + n_2 + 1, \cdots, n_1 + n_2 + n_3$.

Suppose we test whether the gene i is differentially expressed or not. When we have a paired data set, the focus of attention is the mean of all possible differences, $\overline{D_i} = \sum_{j=1}^{n_1} D_{ij} / n_1 = \sum_{j=1}^{n_1} (X_{ij} - Y_{ij}) / n_1$. In two independent sample data sets, the sample mean difference, $\overline{X_i} - \overline{Y_i}$, can be a measure of interests. Note that these two statistics, $\overline{D_i}$ and $\overline{X_i} - \overline{Y_i}$, are the basis of a paired t-test and a two sample t-test, respectively. We combine these two statistics to $w_1 \overline{D_i} + w_2 (\overline{X_i} - \overline{Y_i})$ with weights of $w_1 = n_1$ and w_2, a harmonic mean of n_2 and n_3.

Let ${s_D}^2$, ${s_X}^2$ and ${s_Y}^2$ be sample variances based on $\{D_{ij}\}_{j=1}^{n_1}$, $\{X_{ij}\}_{j=n_1+1}^{n_1+n_2}$ and $\{Y_{ij}\}_{j=n_1+n_2+1}^{n_1+n_2+n_3}$, respectively. Under the assumption that the i[th] gene is not differentially expressed, the null distribution of the following t_3-statistic can be either assumed to approximately follow N(0,1) as n_1, n_2 and n_3 get large or it can be estimated using permutation or bootstrap methods.

$$t_3 = \frac{w_1 \overline{D_i} + w_2 (\overline{X_i} - \overline{Y_i})}{\sqrt{{w_1}^2 \frac{{s_D}^2}{n_1} + {w_2}^2 (\frac{{s_X}^2}{n_2} + \frac{{s_Y}^2}{n_3})}}$$

Sometimes we need to rank genes to prioritize the development of biomarkers for the population screening of cancer. In that case the order of magnitude of absolute t_3 is more important than p-values based on its distribution.

In our example, we first plan to choose differentially expressed genes between normal and colorectal tumor samples using adjusted p-values by Ge et al's max procedure[4]. We found that more than 1000 genes were selected even with very small adjusted p-value like 0.00001. Therefore, the information of absolute t_3 values is sufficient to select differentially expressed genes.

A diagonal linear discriminant analysis (DLDA) and a diagonal quadratic discriminant analysis (DQDA)[5] are used to classify tumors based on the differentially selected genes.

4. Results

We split the mixed data set into a test set and a training set using by a following way.

Test set : two independent samples, 32 tumors and 19 normal tissues.

Training set : among 36 pairs, first 30 pairs, 3 tumor samples from the next 3 pairs (by dropping the paired data on normal tissues) and 3 normal samples from the rest pairs (by dropping the paired data on tumors)

When we have a mixed data set like the training set above to analyze, we usually drop the two independent samples and apply the paired t-test only to 30 paired samples, which loses efficiency by not utilizing all the available information. But, by adopting a new t_3 test statistic, all the samples in a form of mixed set can be used. In a training set, we calculate absolute t_3 values of each gene and the orders of significance of differentially expressed genes are decided. In classifying 51 test set objects, DQDA and DLDA are employed and the genes that are involved to work for a classifier are decided by their orders of significance. Error rates of classification are shown in Table 1 that the top 8 genes are sufficient for DQDA to yield the 0% test error and DLDA needs more.

Table 1 : The classification error rates based on top n genes of t3 statistic

Number of genes(n)	1	2	3	4	5	6	7	8-12	13	14	15 or more
DLDA	0.059	0.039	0.020	0.020	0.020	0	0.020	0	0.020	0.020	0
DQDA	0.059	0.039	0.020	0.020	0	0	0.020	0	0	0	0

5. Discussion

Dudoit *et al.* [5] showed that the DLDA yielded the lowest test error rate even with its simplicity when they compared several discriminant methods including DQDA using lymphoma, leukemia and NCI 60 data sets. But, in our colorectal data set of tumors and normal tissues, which is more heterogeneous than the data sets used by Dudoit *et al.* [5], we found that the DQDA is more efficient than the DLDA.

In the respect of the error rate, we cannot compare the efficiency between t_3 test and other tests based on t-statistics. However, in the matter of data utilization, t_3 test is superior to others.

References

1. Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., and Levine A.J.(1999), Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Nati Acad Sci USA, *96*, 6745-6750.
2. Bo T.H. and Jonassen I.(2002), New feature subset selection procedures for classification of expression profiles, Genome Biology, *3(4)*, research 0017.1-research 0017.11.
3. Yang Y.H., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., Speed T.P.(2002), Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., *30(4):* e15.
4. Ge Y., Dudoit, S. and Speed T. (2003), Resampling-based multiple testing for microarray data analysis. TEST, *12(1),* 1-44.
5. Dudoit S., Fridland J., Speed T.P.(2002), Comparison of discrimination methods for classification of tumors using gene expression data, Journal of American Statistician Associations, *97(457),* 77-87.