Detecting differentially expressed genes for the colorectal cancer: Combining paired data set with two independent data sets

Byung Soo Kim[1], Sunho Lee[2], Inyoung Kim[3], Sang-cheol Kim[3], Sun Young Rha[3], Hyun Cheol Chung[3]

1. Dept. of Applied Statistics, Yonsei University, Seoul, 120-749. bskim@yonsei.ac.kr. 2. Dept. of Applied Mathematics, Sejong University, Seoul, 143-747. 3. Cancer Metastasis Research Center, College of Medicine, Yonsei University, Seoul, 120-752, S.Korea.

1. Aim of the Study

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.

2. Experiment, Data and Statistical Issues

2.1 Experiment

We collected cancer and normal tissues from 58 colorectal cancer patients during the operation and snap froze the tissues at –70C. We attempted to extract total RNAs from tumour and normal tissues from 58 patients. From each of 20 patients we have RNA samples both for tumour and normal tissues. From 16 (22) patients RNA samples for normal (tumour) tissues only were available. Thus, we have a matched pair sample of size 20 and two independent samples of sizes 16 and 22. We conducted a cDNA microarray experiment using a common reference design with 17K human cDNA microarrays. We pooled eleven cancer cell lines and used it for the common reference. We used $M \equiv \log(R/G)$ for the evaluation of relative intensity, where R and G in (R, G) represent the cy5 and cy3 fluorescent intensities, respectively. We normalized the intensity ratio using within-print tip group, intensity dependent normalization following Yang *et al.* (2002).

2.2 Data

For 20 patients with a matched pair data set $\{(X_i, Y_i)\}_{i=1}^{n}$, where X_i and Y_i represent M values for the (Ref, Normal tissue) and (Ref, Tumour) hybridizations, respectively, for the i-th patient. For 16 patients we observed M values only for (Ref, Normal tissue) hybridizations and this M value is denoted by U, which has the same marginal distribution with X ($=X_1$). We also have 22 M values only for (Ref, Tumour) hybridizations, which is represented by V. V is identically distributed with Y ($=Y_1$). Thus, we have following three data types.

Table 1: Three data types of the experiment

Hybridization		No. of Cases
(Ref, Normal tissue)	(Ref, Tumor)	
X	Y	20 (n1)
U	missing	16 (n2)
missing	V	22 (n3)

2.3 Statistical Issues

As a means of utilizing the whole data sets we first use the matched pair data set as a training set from which we detect a set of DE genes between the normal tissue and the tumour.Then we use two independent data sets $\{U_i\}_{i=1}^{16}$ and $\{V_j\}_{j=1}^{22}$ for the test set for validating the chosen set of DE genes. However, we still raise the following two statistical issues for the analysis.

(1) As an initial attempt of employing a multivariate method we propose using Hotelling's T^2 statistic for the detection of a set of DE genes.

(2) We propose a t-based statistic, say t_3, which combines three data types for the detection of DE genes.

3. Results

3.1 Detecting DE genes based on the matched pair data set

We employed the following three procedures for detecting a set of DE genes from the matched pair sample of size 20.

(1) Paired t test and Dudoit *et al.*'s max T procedure for controlling the family-wise error rate (FWER) (Dudoit *et al.*, 2002b).

(2) Tusher *et al.*'s SAM procedure. (Tusher *et al.*, 2001)

(3) Lönnstedt and Speed's empirical Bayes procedure using B statistics (Lönnstedt and Speed, 2002)

Even for the FWER of 0.01 using Procedure (1) we could detect more than 700 genes for the differential expression, which far exceeds the number of candidate genes for the biomarker development. Using each of three procedures we could detect the top 139 genes. [Originally we would like to have top 100 genes. We found that among the top 130 genes of Procedure (1) the last 40 genes have the same adjusted p-values. We also found that 139 was the closest number of DE genes that SAM could detect.] These three procedures reasonably coincide with each other as Table 2 shows.

Table 2: The number of DE genes detected by each procedure

	paired t + max T	SAM	B statistic
paired t + max T	139	110	124
SAM	110	139	123
B statistic	124	123	139

3.2 Classifying the test set: validating the set of DE genes

We restricted ourselves to the top 50 genes for the classification of the test set which comprised 16 normal and 22 tumour specimens. Using these 50 genes we employed the diagonal quadratic discriminant analysis (DQDA) for the classification. We found that only the top 5 genes were required for achieving a 0% test error.

Dudoit *et al.* (2002b) showed that the diagonal linear discrimnant analysis (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. We found in this colon data set that DQDA needed only the top 5 genes, whereas DLDA required the top 7 genes for achieving 0% test error. Our colon data set which consists of tumours and normal tissues is more heterogeneous than the data sets used by Dudoit *et al.* (2002b). This heterogeneity motivated us to use different variances for two groups in the discriminant analysis which is DQDA, and it turned out to be more efficient than DLDA.

3.3 Hotelling's T^2 Statistic

We computed Hotelling's T^2 statistic by paring two genes in all possible ways from the training set and obtained the top 25 pairs in the order of the magnitude. It is interesting to note that this list of 25 pairs has less than 40% overlap with the top 50 gene list of the univariate t statistic. Hotelling's T^2 statistic for a pair of genes is a function of several parameters including the correlation coefficient (ρ). It is a decreasing function of ρ, when other things are equal. Therefore, T^2 statistic can detect some of genes that are not detected by the univariate t test such as gene 1' in Table 3, but has high correlation with a gene of very large t value. We employed DQDA for classifying the test set based on the list of top 25 pairs and found that the top 3 pairs in Table 3 were sufficient to yield the 0% test error.

Table 3: Top 3 pairs of genes selected by Hotelling's T^2 statistic

Gene	Hotelling's T^2	univariate t	correlation
1	1403.9	18.6	-0.82
1'		3.5	
2	1400.6	17.2	0.78
2'		-7.6	
3	1174.1	-18.8	0.61
3'		11.4	

Figure 1: Scatter plots of two pairs of genes in Table 3

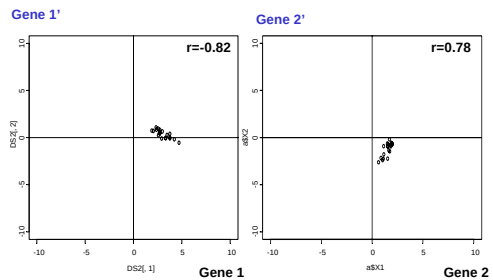

Table 3 shows Hotelling's T^2 statistic and its related statistics for the top 3 pairs of genes. Figure 1 displays that genes 1' and 2' have small t values but have high correlations with genes of large t values. Genes 1, 2, and 3 in Table 3 are contained in the top 5 genes of Sec 3.2 found by DQDA and the univariate t statistic. This result strongly indicates that the multivariate approach warrants further research in the microarray analysis.

3.4 Proposing a t-based statistics, t_3

For detecting a DE gene on the basis of a univariate t statistic we may propose the following t_3 statistic as a means of pooling matched pair and two independent data sets. Following the notations in Table 1 we define $D_i = X_i - Y_i$, i=1, …, n_1. Let S_D^2, S_U^2 and S_V^2 are sample variances based on $\{D_i\}_{i=1}^{n1}$, $\{U_j\}_{j=1}^{n2}$, and $\{V_t\}_{t=1}^{n3}$, respectively. Then under the no DE gene hypothesis the following t_3 statistic has an approximately N(0,1) distribution.

$$t_3 = \frac{n_1 \overline{D} + n_h(\overline{U} - \overline{V})}{\sqrt{n_1 S_D^2 + n_h^2 [\frac{1}{n_2} S_U^2 + \frac{1}{n_3} S_V^2]}} \quad ,$$

where the 'bar' notation denotes the sample mean and n_h is the harmonic mean of n_2 and n_3.

4. Discussions

It is interesting to note that DQDA works better than DLDA for this tumour versus normal tissue data. It is quite desirable to develop a "stepwise classification" to further narrow down the DE genes that achieve the same test error rate with the top 5 or top 3 pairs.

5. References

Dudoit S, Fridlyand J , Speed TP. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Amer. Statist. Assoc*. 97(457):77-87.

Dudoit S, Yang YH, Callow MJ, Speed TP (2002b).Statistical methods for identifying differentially expressed genes in r eplicated cDNA microarray experiments, *Statistica Sinica* 2(1):111-139.

Lönnstedt I, Speed TP. (2002). Replicated microarray data. *Statistica Sinica*, 12(1):31-46.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*., 30(4):e15.

Tusher V, Tibshirani R, Chu, G. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci*., 98:5116-5121.