Statistical Analysis for Feature Subset Selection Procedures.

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

[1] Cancer Metastasis Research Center, Yonsei University College of Medicine, Seoul, Korea
[2] Department of Applied Mathematics, Sejong University, Seoul, Korea
[3] Department of Applied Statistics, Yonsei University, Seoul, Korea

[*]To whom correspondence should be addressed. E-mail: bskim@yonsei.ac.kr

Abstract

In this paper, we propose using Hotelling's T2 statistic for the detection of a set of a set of differentially expressed (DE) genes in colorectal cancer based on its gene expression level in tumor tissues compared with those in normal tissues and to evaluate its predictivity which let us rank genes for the development of biomarkers for population screening of colorectal cancer. We compared the prediction rate based on the DE genes selected by Hotelling's T2 statistic and univariate t statistic using various prediction methods, a regulized discrimination analysis and a support vector machine. The result shows that the prediction rate based on T2 is better than that of univatiate t. This implies that it may not be sufficient to look at each gene in a separate universe and that evaluating combinations of genes reveals interesting information that will not be discovered otherwise.

Introduction

This paper is a part of ongoing research project to identify a set of differentially expressed (DE) genes in colorectal cancer based on its gene expression level in tumor tissues compared with those in normal tissues and to evaluate its predictivity which let us rank genes for the development of biomarkers for population screening of colorectal cancer.

One of the standard methods for selecting feature subset is detecting DE genes according to univariate t statistic, which is the weighted distance between tumor and normal tissues on each gene, and computing the prediction rate on a test set with selected genes using favor prediction methods, e.g., discrimination analysis, which are the multivariate tools. Even though univariate t statistic provides simple and speedy detection of DE genes, it treats each gene independently, which is not necessarily true, and eventually provides inefficient selection. In addition, it is not coherent to carry out multivariate prediction analysis with genes selected by the one dimensional method, t statistic, which may result in including irrelevant genes. Hence, we propose multivaiate Hotelling's T2 statistic for the detection of a set of DE genes.

In the paper, we compare the prediction rate based on the DE genes selected by Hotelling's T2 and univariate t statistic using several prediction methods, a regulized discrimination analysis (Friedman, 1989), support vector machine (Cristianini and Shawe-taylor, 2000), to investigate which one is more efficient.

Materials and Methods

Colorectal tissue samples and RNA preparation

We collected cancer and normal tissues from 87 colorectal cancer patients during the operation at the Severance Hospital, Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Korea. Tissue samples were immediately frozen into liquid nitrogen and stored at -80℃ until further use. We attempt to extract total RNAs from tumor and normal tissues from all patients using Trizol reagent (Invitrogen, USA) according to the manufacturer's protocol. Then, extracted RNA was purified before probe preparation using an RNeasy kit (Qiagen, Germany) based on the supplier's manuals. The quantity and quality of RNA were evaluated using a Gene Spec III (Hitachi, Japan), and a Gel Documentation-Photo System (Vilber Lourmat, France), respectively.

Microarray experiment

We conducted a cDNA microarray experiment using a common reference design with 17K human cDNA microarrays (GT-CMRC, Korea) including 9K CGAP clones (Research Genetics, USA). For the common reference, we used pooled RNA of eleven cancer cell lines of various origins, which is currently using as a standard reference in Cancer Metastasis Research Center (CMRC), Yonsei University College of Medicine, Korea. Probe preparation and microarray hybridization were done following the standard protocol of Yonsei CMRC. Following hybridization, array was scanned using a GenePix 4000B(Axon Ins., USA) and the images were analyzed using GenePix Pro 3.0(Axon Ins., USA).

Data set

We have had 133 microarray data which comprised of 36 paired data set with tumor and normal tissues from one patient, 32 tumor tissues only, and 19 normal tissues only. 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 of Yang et al.(2002) is adopted as a normalization for log intensity ratio, M=log(R/G), for the evaluation of relative intensity, where, R and G in (R,G) represent the cy5 and cy3 flurescent intensities, respectively. Finally, we obtained 2850 genes expressions. As a means of utilizing the whole data sets we first use the matched pair set as a training set from which we detect a set of DE genes between the normal tissues and the tumour.

Results

Detecting DE genes based on the matched pair data set using univariate t and Hotelling's T2

Univariate t Statistic

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) and (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 1000 genes for the differential expression, which far exceeds the number of candidate genes for the biomarker development. Since the number of candicate genes are still quite large, we first consider only top 100 genes. These three procedures reasonably coincide with each other as Table 1 shows.

Table1: the number of DE genes detected by each procedure using univariate t

	Paired T & maxT	SAM	B Statistic
Paired t+ Max T	100	74	83
SAM	74	100	85
B Statistic	83	85	100

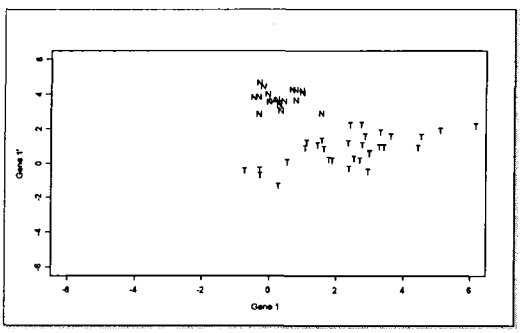
Hotelling's T^2 Statistic

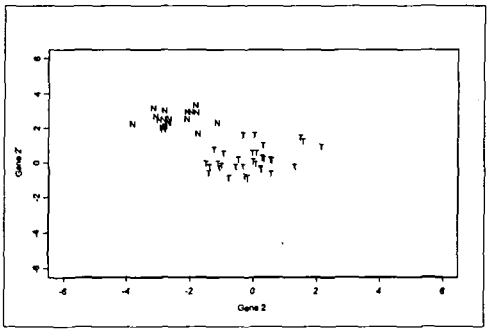
We also computed Hotelling's T^2 statistic by paring two genes in all possible ways (the number of 12850 choose 2) from the training set and considered the top 25 pairs in the order of the magnitude. It is interesting to note that this list of 25 pairs has less than 40% overlap with the top 50 gene list 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 (r). It is a decreasing function of r, when other things are equal. Therefore, Hotelling's T^2 statistic can detect some of genes that are not detected by the univariate t test such as gene 1 in Table 2, but has high correlation with a gene of very large t value. The scatter plots of two pairs of genes are in the figure 1.

Table 2: Top 2 pairs of genes selected by Hotelling's T2 statistic

Gene	Hotelling's T2	Univariate t	Correlation.
G1	1372.29	7.3918	0.749
G1'		-18.492	
G2	1316.21	10.6948	0.6115
G2'		-20.889	

Figure1: Scatter plot of two pairs of genes in Table 2. N and T are the notations of normal and tumor samples, respectively.





Classifying the test set: validating the set of DE genes

Since 50 is the maximum number to do confirmatory experiment, we restricted ourselves to the top 50 genes selected by univariate t and to the top 25 pairs selected by Hotelling's T2 for the classification of the test set which comprised 19 normal and 42 tumor specimens.

Using 50 genes and 25 pairs selected by univariate t and Hotelling's T2, respectively, we employed the regulized discrimination analysis (RDA) to compromise between the diagonal linear discrimination analysis (DLDA) and the diagonal quadratic discriminant analysis (DQDA) which allows one to shrink the separate covariate of DQDA toward a common covariate as in DLDA, as well as a support vector machine (SVM) with various kernel functions for the classification.

We found that only the top 5 genes selected by univariate t statistic were required for achieving a 0% test error using RDA. We also investigated the result by RDA is more shrinked to DQDA for our data set. The test error by this method is the same as that by DQDA. We also noticed that the DQDA works better than DLDA for this tumors versus normal tissues samples. This result is different from Dudoit *et al.* (2002b) which showed that the diagonal linear discriminant analysis (DLDA) yielded the lowest test error rate even with its simplicity when they compared several discriminiant methods including DQDA using lymphoma, leukemia and NCI 60 data sets. But our colon dataset which consists of normal and tumours tissues is more heterogeneous than the data seta used by Dudoit et al. (2002). This heterogeneity motivated us to use different variances for two groups in the discriminant analysis. On the other hand, the top one pair selected by Hotelling's T2 statistic is enough to obtain 0% test error using RDA. Unlike univariate t statistic, the result by RDA is more shrinked to DLDA for our data set.

The same improvement is observed even when we use SVM with various kernels. The prediction rates using RDA and SVM along with DLDA and DQDA, using the top 50 genes which were selected by univariate t and Hotelling's T2, are listed in table 3 and table4, respectively. For SVM, we consider the linear kernel(SVML), quadratic kernel(SVMQ) and gaussian kernel(SVMG).

Table3: the prediction rate by various prediction methods using top 50 genes which were selected by univariate t. The n of the first column means that the test is done with the top n genes.

N	DLDA	DQDA	RDA	SVML	SVMQ	SVMG
1	0.94	0.94	0.94	0.37	0.52	0.94
2	0.98	0.98	0.98	0.98	1	0.98
3	0.96	0.98	0.98	0.98	1	0.98

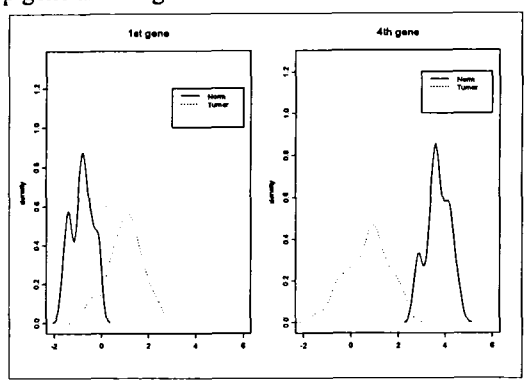
4	0.98	0.98	0.98	1	1	1
5	0.96	1	1	1	1	1
6	0.96	1	1	1	1	1
7	0.96	1	1	1	1	1
8	0.96	1	1	1	1	1
9	0.98	1	1	1	1	1
>10	1	1	1	1	1	1

Table4: the prediction rates by various prediction methods using top 25 pairs genes which were selected by Hotelling's T2.

	DLDA	DQDA	RDA	SVML	SVMQ	SVMG
1	0.84	0.86	0.84	0.84	0.84	0.84
2	1	0.98	1	1	1	0.98
3	1	0.98	1	1	1	0.98
4	1	1	1	1	1	0.98
5	1	1	1	1	1	0.98
6	1	1	1	1	1	0.98
7	1	1	1	1	1	0.98
8	1	1	1	1	1	0.96
9	1	1	1	1	1	0.94
>10	1	1	1	1	1	0.94

Furthermore, we also investigated that 0% test error was obtained using only one gene which is ranked 4th by univariate t statistic. When we compare this gene with top ranked gene, it is shown that a distance between tumor and normal tissues on top ranked genes is larger than that of 4th gene, but the gene expression distribution of tumor and normal tissues have overlapped part on a top gene unlike 4th gene as figure 2 shows. We also noticed that this 4th gene is one of the top pair genes by Hotelling's T2.

Figure2: the probability density function of gene expression between normal and tumours for top gene and 4th gene



The result shows us that the prediction rate in T2 seems to be better than that of t. This implies that it may not be sufficient to look at each gene in a separate universe and that evaluating combinations of genes reveals interesting information that will not be discovered otherwise. One of the rationales on these observations is that tumor and normal tissues may be well separated in a higher dimension even though the projections on each gene coordinate are highly overlapped. This kind of pairs could not be selected in standard method since each gene may have a small absolute t value. The other rationale is that it is more coherent to carry out multivariate prediction method with genes selected by multidimensional method, T2 so that we achieve the prediction rate 1 with fewer genes. Therefore, this result strongly indicates that the multivariate approach warrants further research in the microarray analysis.

Discussion

It is interesting note that DQDA works better than DLDA for this tumour versus normal tissues data for univariate t statistic on contrast of the case of Hotelling's T2.

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 1 pair.

The concept of T2 can be generalized to higher dimensional test T2(n), which can be defined as between normal and tumor tissues in n-gene coordinates. We select n-genes pairs according to T2(n) statistic as we do in T2 statistic. T2(n) will be more efficient when the gene expression distribution of tumor and normal tissues have highly overlapped in lower dimensions. If we increase the dimension simply, there could be higher chance to select irrelevant genes. Hence, it is interesting to know the optional n. In our example, one pair selected by T2 statistic is enough to achieve the prediction rate 1. Thus, we can say the optional n in our example is 2.

Acknowledgements

The research of Byung Soo Kim 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). The research of Hyun Cheol Chung was supported by the Korea Science and Engineering Fund through the Cancer Metastasis Research Center at Yonsei University.

References

[1] S. Dudoit, J. Fridlyland, and T. P., Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Amer. Statist. Assoc.*, 97(457), 2002, 77-87.

[2] S. Duodit, Y. J. Yang, M. J. Callow and T.P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray data, *Statisical Sinica*, 2(1), 2002, 111-139.

[3] J. Friedman, Regulized discriminant analtsis, Journal of the American Statistical Association, 84, 1989, 24-26.

[4] I. Lonnstedt and T. P. Speed, Replicated microarray data, *Statistical Sinica*, 12(1), 2002, 31-46.

[5] N. Cristianini and J. Shawe-Taylor, An Introductiion to Support Vector Machines, Cambride, UK, 2000.

[6] V. Tusher, R. Tibshirani, G. Chu, Significace analysis of microarrays applicated to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci.*, 2001, 98, 5116-5121.

[7] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, T. P. Speed, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Res.*, 2002, 30(4), e15.

[8] V. Tusher, R. Tibshirani, G. Chu, Significace analysis of microarrays applicated to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci.*, 2001, 98, 5116-5121.