Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer

Byung Soo Kim1,*, Inyoung Kim2, Sunho Lee4, Sangcheol Kim2,3, Sun Young Rha2,3 and Hyun Cheol Chung2,3

1Department of Applied Statistics, 2Cancer Metastasis Research Center, College of Medicine and 3Brain Korea 21 Project for Medical Science, College of Medicine, Yonsei University, Seoul, South Korea and 4Department of Applied Mathematics, Sejong University, Seoul, South Korea

Received on June 29, 2004; revised on August 28, 2004; accepted on September 12, 2004

Advance Access publication September 16, 2004

ABSTRACT

Motivation: It is a common practice in cancer microarray experiments that a normal tissue is collected from the same individual from whom the tumor tissue was taken. The indirect design is usually adopted for the experiment that uses a common reference RNA hybridized both to normal and tumor tissues. However, it is often the case that the test material is not large enough for the experimenter to extract enough RNA to conduct the microarray experiment. Hence, collecting n cases does not necessarily end up with a matched pair sample of size n. Instead we usually have a matched pair sample of size n1, and two independent samples of sizes n2 and n3, respectively, for ‘reference versus normal tissue only’ and ‘reference versus tumor tissue only’ hybridizations (n = n1 + n2 + n3). Standard statistical methods need to be modified and new statistical procedures are developed for analyzing this mixed dataset.

Results: We propose a new test statistic, $T^2$, as a means of combining all the information in the mixed dataset for detecting differentially expressed (DE) genes between normal and tumor tissues. We employed the extended receiver operating characteristic approach to the mixed dataset. We devised a measure of disagreement between a RT–PCR experiment and a microarray experiment. Hotelling’s $T^2$ statistic is employed to detect a set of DE genes and its prediction rate is compared with the prediction rate of a univariate procedure. We observe that Hotelling’s $T^2$ statistic detects DE genes more efficiently than a univariate procedure and that further research is warranted for ‘reference versus normal tissue’ and ‘reference versus tumor tissue’ comparisons.

Contact: bskim@yonsei.ac.kr

1 INTRODUCTION

Microarray technology, with its potential to quantitatively measure the expression levels of thousands of genes simultaneously, holds the promise of becoming a valuable tool in cancer research and clinical diagnostics. Two technologies are widely used, the cDNA microarray and the oligonucleotide array, which differ with respect to the types of nucleic acid probes arrayed for the interrogation of labeled RNA specimens. In the present study, we focus on the cDNA microarray, which uses a dual-label system in which two RNA specimens are separately reverse transcribed, labeled, mixed and hybridized together to each array.

In the cancer microarray experiment based on n patients, the experimenter often wants to compare the tumor tissue with the normal tissue within the same individual using a common reference RNA. This design is often referred to as a reference design or an indirect design. Ideally, this experiment produces n pairs of microarray data, where each pair consists of two sets of microarray data resulting from reference versus normal tissue and reference versus tumor tissue hybridizations. However, for certain individuals either normal tissue or tumor tissue is not large enough for the experimenter to extract enough RNA for conducting the microarray experiment, hence there are missing values in the normal and tumor tissue data. Practically, we have n1 pairs of complete observations, n2 ‘normal only’ and n3 ‘tumor only’ data for the microarray experiment with n patients, where n = n1 + n2 + n3. We refer to this dataset as a mixed dataset, as it contains a mixture of fully observed and partially observed pair data. This mixed dataset was actually observed in the microarray experiment based on human tissues, where human tissues were obtained during the surgical operations of cancer patients. There are two major points that the reference design gains in this type of microarray experiment. First, it permits the parallel comparison of differentially expressed (DE) genes along different cancers. Second, it allows us to utilize the partially observed pair data obtained from n2 and n3 patients. If we had employed the direct design where
both tumor and normal tissues were put on a single array, no information would have been obtained for the partially observed pair data.

The aim of this study is to develop a statistical method for the mixed dataset. We conducted cDNA microarray experiment based on 87 colorectal cancer patients using a reference design with cDNA microarrays containing ~17,000 human genes. The primary objective of the microarray experiment based on colorectal cancers was to ranking genes for the development of a biomarker that could be used for the population screening and detecting DE genes for various subtypes of colorectal cancer as well, e.g. colon cancer versus rectum cancer, stages (B, C and D) and carcinoembryonic antigen (CEA) ≥ 5 versus CEA < 5. The clinicians use CEA for monitoring the progress of the colorectal cancer after the initial treatment, and the threshold value is usually set to 5.

The success of microarray technology in cancer research depends on the development of statistical methods for detecting DE genes among different tissue types. The most widely used statistical methods for detecting DE genes are basically univariate procedures often coupled with adjustment of p-values or a similar concept due to multiple tests (Tusher et al., 2001; Dudoit et al., 2002; Lönnstedt and Speed, 2002). These univariate methods disregard the multidimensional structure of microarray data. Multivariate approaches are needed that can utilize the correlated structure of the microarray data and therefore capture the hidden information in gene interactions. Li et al. (2001) proposed a genetic algorithm combined with the k-nearest neighbor method to detect DE genes that might have a potential to jointly discriminate different tissues, e.g. normal versus tumor. However, it still remains to be seen whether the genetic algorithm can detect DE genes in multivariate dimension. Szabo et al. (2003) suggested a random search algorithm after defining a computationally tractable distance in the p-dimensional gene space. It is interesting to note that the set of DE genes detected by a univariate procedure, for example, the t-test and the set of DE genes selected by these two multivariate procedures have ≤ 50% in common for the biological replicates. We propose in this study using Hotelling’s T^2 statistic for detecting DE genes. As an initial attempt to employ Hotelling’s T^2 statistic as a means of multivariate approach in the microarray data, we restrict ourselves to a random vector of length two primarily due to the computational limit.

Simon and Dobbin (2003), characterized the objectives of many DNA microarray experiments as class comparison, class prediction or class discovery. The detection of DE genes between two groups, e.g. between tumors that respond to chemotherapy and those that do not (Rosenwald et al., 2002), between normal and tumor tissues (Alon et al., 1999) and among various subtypes of a tumor (Hedenfalk et al., 2001), comprises a class comparison problem. Classifying a new specimen to the known subtype of a cancer (van de Vijver et al., 2002) and developing a molecular prognostic predictor model (Rosenwald et al., 2002) would result in a class prediction problem. Discovering a new subtype of a cancer based on the molecular profile (Alizadeh et al., 2000) would belong to the class discovery problem. In addition to these objectives, ranking candidate genes for the development of a biomarker that can be used for the population screening of a cancer should serve the purpose of the microarray experiment. Pepe et al. (2003) elegantly employed the receiver operating characteristic (ROC) approach for this purpose based on two independent samples that consist of normal and tumor tissues. We employ the extension of Pepe et al.’s approach to the mixed dataset and rank candidate genes involved in the development of a biomarker for the population screening of the colorectal cancer. The extension procedure of Pepe et al.’s approach will be reported in a separate communication.

We first show that three commonly used univariate procedures produce more or less the same set of DE genes between normal and tumor tissues. For the validation of statistical procedures, either univariate or multivariate, used in this study for the detection of DE genes between normal and tumor tissues, the mixed dataset provides a natural way of splitting sample. We use the n_1 paired data for the training set, and the pool of n_2 ‘normal only’ and n_3 ‘tumor only’ data for the test set. We employed several classifiers including linear and quadratic discriminant analyses, and support vector machine (SVM) for classifying the test set based on the DE genes detected from the training set. Using this split sample approach, we show that Hotelling’s T^2 statistic detects DE genes more efficiently than a commonly used univariate procedure. Now, as a means of combining all the information in the mixed dataset we propose a t-based statistic, t_3, for the detection of DE genes between normal and tumor tissues. We develop a measure of disagreement between a RT–PCR experiment and a microarray experiment. For the subtype analyses, we employ the split sample approach 10 times repeatedly for the validation of DE genes in each class of subtypes. We use the two-sample t-statistic coupled with a P-value adjustment, or the similar concept due to multiple tests and SVM to calculate the prediction rate. We could improve the test error of ‘CEA ≥ 5’ with the help of the regression-based gene selection by taking advantage of the continuous scale of CEA.

In Section 2, we describe patient and experiment data, and microarray data pre-processing including normalization and missing value imputations. Section 3 presents three univariate procedures, Hotelling’s T^2 statistic, t_3 statistic for the detection of DE genes, several classifiers, the ROC approach for ranking genes for the biomarker development, regression-based gene selection for the CEA variable and a measure of disagreement between a RT–PCR assay and a microarray assay. Technical details are discussed in the Appendix section.
Section 4 shows the results and Section 5 concludes with a discussion.

2 EXPERIMENT AND DATA PRE-PROCESSING

Cancer and normal tissues were obtained during the surgical operations from 87 colorectal cancer patients at Severance Hospital (Yonsei Cancer Center, Yonsei University College of Medicine, Seoul, Korea) from May to December 2002. We conducted a cDNA microarray experiment using a common reference design with cDNA microarrays containing ~17 000 human genes. We pooled 11 cancer cell lines and used it for the common reference. These 11 cancer cell lines are as follows: AGS, MDA-MB-231, HCT-116, SK-HeP-1, A549, HL-60, MOLT-4, HeLa, HT-1080, Caki-2 and U87MG (American Type Culture Collection). The fresh specimens of cancer and normal tissues obtained from the colorectal cancer patients during surgery were snap-frozen in liquid nitrogen immediately after the resection and stored at -70°C until further use. Clinical characteristics of the patients are provided in Table 1.

We originally attempted to extract total RNAs from the tumor and normal tissues of 87 patients. We obtained RNA specimens both for tumor and normal tissues from 36 patients. However, RNA specimens for normal tissues alone were available from 19 patients. Also, RNA specimens for tumor tissues alone were obtained from the other 32 patients. Thus, we have a matched pair sample\(^2\) of size 36 and two independent samples of sizes 19 and 32. In terms of notations provided in Section 1: \(n_1 = 36, n_2 = 19\) and \(n_3 = 32\). These tissues were taken by a single surgeon from 87 patients, and there was no specific clinical or biological meaning on the possibly different characteristic among these three subgroups. Therefore, we assume that these three subgroups are independent samples from a single population. After total RNAs were extracted from fresh frozen tissues, 50 \(\mu g\) of purified RNAs were labeled and hybridized to cDNA microarrays based on the protocol established in Cancer Metastasis Research Center (Yonsei University, Korea) (Park et al., 2004).

We used \(M = \log_2(R/G)\) for the evaluation of relative intensity, where \(R\) and \(G\) represent the cy5 and cy3 fluorescent intensities, respectively.

Quantitative real-time RT–PCR was performed with six selected genes using a Rotor Gene 2072D real-time PCR machine (Corbett Research, Australia) in accordance with the manufacturer’s instructions. We used SYBR Green (Quiagen, CA) for the labeling. The amplified fluorescence signal was measured and the level of transcript for each specimen was calculated based on the standard curve. The standard curve was drawn by plotting the measured threshold cycle versus the arbitrary unit of copies/reaction according to the amount of serially diluted standard RNA. The threshold cycle (Ct) value was determined as the cycle number at which the fluorescence exceeded the threshold value.

Let \(X\) and \(Y\) denote the log-fluorescent intensity ratios of reference versus normal and reference versus tumor hybridizations, respectively. Let \(U\) and \(V\) be independent copies of \(X\) and \(Y\), respectively. Then, we may observe three data types represented as follows:

\[
\left(\frac{(X_j, Y_j)}{n_1}\right)_{j=1}^{n_1}, \left\{U_k\right\}_{k=1}^{n_2} \text{ and } \left\{V_l\right\}_{l=1}^{n_3}.
\]

We first define no missing proportion (NMP) of a gene as the proportion of valid observations out of the total number of arrays. For example, if a gene has valid observations for 32 out of 40 arrays, its NMP is 0.8. We pre-processed the data as follows:

1. We normalized the log-intensity ratio, \(\log_2(R/G)\), using within-print tip group, intensity-dependent normalization following Yang et al. (2002).
2. We used 0.8 for the cut point of NMP to delete genes containing missing values for >20% of the total number of observations. This filtering procedure yielded 13 859 genes.
3. We employed \(k\)-nearest neighbor \((k = 10)\) method for the imputation of missing values.
4. We averaged values for the multiple spots. The numbers of duplicated,triplicated and quadruplicated spots were 982, 6 and 5, respectively.
5. Finally, we have a dataset represented using a 12 850 \(\times\) 123 matrix, where 12 850 represents the number of genes and 123 stands for the number of microarrays.

We investigated various box plots (data not shown) after the localization (location parameter) normalization, and concluded that it was not necessary to have the scale normalization either between blocks within an array or between arrays.

One way of utilizing all the information in the mixed dataset would be using the matched pair sample of a training set for the detection of DE genes and two independent samples of a test set for the validation of the chosen DE genes. By detecting a set of DE genes in the training set of matched pairs we could control between individual variation. The independence within and between two samples of sizes 19 and 32 would make them the ideal test set.

Table 1. Clinical characteristics of 87 colorectal cancer patients

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Location</th>
<th>CEA</th>
<th>Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
<td>Median</td>
<td>Range</td>
<td>Colon</td>
</tr>
<tr>
<td>46</td>
<td>41</td>
<td>65</td>
<td>28-90</td>
<td>41</td>
</tr>
</tbody>
</table>

All numbers indicate frequencies except for age.

We use ‘sample’ to denote a random sample in statistics to distinguish it from a biological specimen.
3 STATISTICAL METHODS

3.1 The Detection of DE genes between normal and tumor tissues

We employ the following three univariate procedures for the detection of a set of DE genes from the matched pair sample of size 36.

(1) Paired t-test and Dudoit et al.’s maxT procedure for controlling the family-wise error rate (FWER) (Dudoit et al., 2002b; Ge et al., 2003).
(3) Lönnstedt and Speed’s (2002) empirical Bayes procedure using B-statistic.

Dudoit et al. (2002b) employed the FWER for controlling the Type I error and used Westfall and Young’s step down procedure for calculating the adjusted P-value. We have $2^{36}$ permutations of changing the signs of the paired t-statistic for 36 patients from which we can derive the null distribution of the paired t-statistic. We used 100 000 bootstrap samples of size 36, due to computation limit, to derive the null distribution of the paired t-statistic. Tusher et al.’s (2001) 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 as the expectation of the number of false positive genes divided by the number of declared significant genes. FDR is more sensitive in the detection of significant genes (Ge et al., 2003). Lönnstedt and Speed (2002) used empirical Bayes method to derive a Bayes log posterior odds, e.g. B-statistic. The experimenter may consider top 100 genes in terms of B-values in combination with experimental preference.

We computed Hotelling’s $T^2$ statistic by pairing two genes in all possible ways from the training set and considered the top 25 pairs in the order of magnitude. We restricted ourselves to the top 50 genes, since 50 is the maximum number of genes for the experimenter to conduct a confirmatory experiment. Hotelling’s $T^2$ statistic can be expressed as follows:

$$T^2 = \frac{1}{1 - \rho^2}(t_1^2 + t_2^2 - 2\rho t_1 t_2),$$

where $t_1$, $t_2$ and $\rho$ denote t-statistic for the two component genes, and their correlation coefficient, respectively (Mood et al., 1974). Hotelling’s $T^2$ statistic can pick up some of the genes that are not detected by the univariate t-statistic mostly when either $t_1$ or $t_2$ is large, and $\rho |t_1 t_2|$ has a negative sign and $|t_1 t_2|$ is large. To validate sets of DE genes detected by various statistical methods from the training set, we applied several classification methods to the test set including the diagonal linear discriminant analysis (DLDA), the diagonal quadratic discriminant analysis (DQDA) following Dudoit et al. (2002a) and SVM with several kernels, such as linear, quadratic and Gaussian kernels (Christiani and Shawe-Taylor, 2000).

As a means of utilizing all the information of the mixed dataset for the detection of DE genes, we propose using a t-base statistic, $t_3$, as follows:

$$t_3 = \frac{n_1 \bar{D} + n_H (\bar{V} - \bar{U})}{\sqrt{n_1 S_D^2 + n_H^2 (1/n_1 S_U^2 + 1/n_H S_V^2)}}.$$

where $\bar{D} = \frac{1}{n_1} \sum_{j=1}^{n_1} D_j = \frac{1}{n_1^2} \sum_{j=1}^{n_1} (Y_j - X_j)$, $\bar{U} = \frac{1}{n_2} \sum_{k=1}^{n_2} U_k$, $\bar{V} = \frac{1}{n_2^2} \sum_{k=1}^{n_2} V_k$, $S_D^2$, $S_U^2$, and $S_V^2$ are sample variances of $D$, $U$ and $V$, respectively, and $n_H$ is the harmonic mean of $n_1$ and $n_2$. The distributions of $D$, $U$ and $V$ are almost symmetric around their means under the null hypothesis. Thus, small sample sizes such as 20 for $n_1$, $n_2$ and 3 would invoke the central limit theorem to approximate the null distribution of $t_3$ from Equation (2) by $N(0,1)$.

3.2 Ranking candidate genes for the biomarker development

Recently, Pepe et al. (2003) employed the ROC approach to rank candidate genes that can be used for the biomarker development with the ultimate purpose of population screening of a cancer. Identifying DE genes in a cancer can be accomplished using various statistical methods, for example, those described in Section 3.1. If a gene is found to be differentially expressed in the cancer, then by developing a suitable biomarker, the corresponding protein product or an antibody to it can be detected in the blood or urine, which forms the basis for the population screening (Pepe et al., 2001).

When we develop a biomarker that can be used for the population screening of a specific cancer from a microarray experiment, we note the following two points. First, clinical bioassays for some gene products may be too difficult to develop for technical reasons. Thus, we need to have a sizable number of candidate genes with development priorities so that if one gene proves to be useless for biomarker development, we may still explore the next gene for the development. The second point is that the bioassay, once it is developed, is applied to the whole population, and hence the false positive rate should be extremely low. Even a small false positive rate yields a large number of healthy people being subjected to diagnostic procedures that are unnecessary, costly and sometimes invasive. For ranking candidate genes under these two perspectives, we need statistical measures that discriminate between normal and tumor tissues. The measure of choice should focus on the minimization of the false positive probability as well as on the separation of these two distributions. Pepe et al. (2003) provide the rationale of using ROC approach based on two independent samples for this purpose of ranking candidate gene instead of using t- or Mann–Whitney statistics.

We first modified Pepe et al.’s ROC approach of ranking genes for a paired dataset of size $n_1$ and referred to the corresponding ROC value as ROC$_{pair}$. Pepe et al.’s approach was
directly applicable to calculate the ROC value, denoted by 
ROC_{\text{ind}} for two independent samples of sizes \( n_2 \) and \( n_3 \) in the mixed dataset. Then, we averaged these two ROC values to 
derive an overall ROC value, denoted by \( \text{ROC}_{\text{mix}} \), as follows:

\[
\text{ROC}_{\text{mix}}(t_0) = \frac{n_1 \text{ROC}_{\text{pair}}(t_0) + n_H \text{ROC}_{\text{ind}}(t_0)}{n_1 + n_H},
\]

where \( t_0 \) was the false positive probability which, in turn, 
was determined by a threshold value and \( n_H \) was the har-monic mean of \( n_2 \) and \( n_3 \). Details of the modification and the 
extension of the ROC approach will be reported in a separate 
communication.

### 3.3 Prediction of subtypes

The colorectal cancer dataset contains three clinical variables 
of interest, namely location, CEA value and stage. Each of 
these three clinical variables defines a class of subtypes. The primary interest in the subtype analysis is to detect a set of 
DE genes between subtypes of a given clinical variable and to 
validate the chosen set of DE genes for the classification. For example, we are interested in identifying a set of genes that discriminate the colon cancer from the rectum cancer. For the stage variable, we are interested in the pairwise comparison, e.g., B versus C and so on. We use the split sample approach for the validation by randomly dividing the 68 tumors into a training set of 45 tumors and a test set of 23 tumors. For the subtype analyses, we deleted 19 ‘normal only’ cases of 
which the microarray experiment did not provide any gene 
information on tumor tissues.

We employed the two-sample \( t \)-statistic (with unequal vari-
ances) coupled with FWER, FDR and pFDR (Storey, 2002) to 
detect a set of DE genes for each class of subtypes. We used 
SVM to calculate the prediction rate of each subtype based on the 
top 10 genes in terms of \( t \)-statistic. This process of detect-
ing DE genes and calculating the prediction rate was repeated 10 times each for a pair of training and test sets randomly 
generated from the data, and these 10 prediction rates were 
averaged.

So far as CEA is concerned we employed a regression-based 
gene selection to take advantage of its continuous scale and 
 improve the prediction of CEA \( \geq 5 \). Details of this approach are described in the Appendix section.

### 3.4 Assessing accuracy of the microarray gene 
expression against RT–PCR

We first calculated Pearson’s correlation between log-
transformed RT–PCR, denoted by log(RT–PCR), and the 
 microarray gene expression measurement based on 17 mRNA 
 specimens (9 tumor and 8 normal tissues). The scatter dia-
gram of log(RT–PCR) and the microarray gene expression 
measurement is typically shown in Figure 1 that normal tissue 
group was separated from the tumor group and each group has 
a rather large within-group variation. Obviously, Pearson’s
For the construction of \( \text{diff}_{err} \), we consider RT–PCR as the standard and assume that the error rate of microarray-based decision is not less than the error rate of the RT–PCR-based decision. We illustrate the calculation of \( \text{diff}_{err} \) in Figure 1 based on an example dataset.

4 RESULTS

4.1 Three univariate procedures

Three univariate procedures discussed in Section 3.1 reasonably coincide with each other. There were more than 2000 DE genes screened in the \( t + \text{FWER} \) procedure for which the adjusted \( P \)-values were \(<0.01\). We compared the overlap pattern of the top 100 genes selected from each of these three procedures. For the \( B \)-statistic the top 100th gene had 22.95 for the log-posterior odds ratio, which might have had value 0 under the null hypothesis. For SAM procedure, we decreased the delta (of SAM software) until we obtained 100 DE genes for which the FDR was observed to be 0.0031 (0.31%). Three sets of top 100 genes detected by three procedures have 72 genes in common as shown in Figure 2. Thus, for comparing the performance of univariate procedures with Hotelling’s \( T^2 \) statistic we use just one procedure in this study, the \( t \)-statistic.

4.2 Hotelling’s \( T^2 \) statistic

Hotelling’s \( T^2 \) statistic in Equation (1) indicates that it can pick up some of the genes that are not detected by the univariate \( t \)-statistic, such as \( G'_1 \) and \( G'_2 \) listed in the bottom table of Figure 3, but have high correlations with genes of very large \( t \)-values. Figure 3 shows scatter plots of these two gene pairs in the test set, from which we note that the Hotelling’s \( T^2 \) statistic clearly separates the normal tissue from the tumor tissue and its performance is better than a univariate \( t \)-statistic alone.

4.3 Classifying the test set

We considered top 50 genes selected by the univariate \( t \)-statistic and top 25 pairs chosen by Hotelling’s \( T^2 \) statistic for the classification of the test set that comprises 19 normal and 32 tumor specimens. We increased the number of genes one by one in the classifiers used beginning with the top gene. We observed that 0% test error was achieved with a few genes as shown in Table 2. We further noted that 0% test error was achieved using only one gene, which ranked fourth using the univariate \( t \)-statistic. However, this gene, denoted by \( G_1 \) in the bottom of the table of Figure 3, belongs to the top ranked gene pair of Hotelling’s \( T^2 \) statistic. When we compare the gene expression distribution of the top ranked gene and the fourth ranked gene selected by the \( t \)-statistic, we note that the fourth gene has better separated gene expression distributions between normal versus tumor than the top ranked gene. This aspect is shown in Figure 4. Furthermore, when we look up the upper diagram in Figure 3, we note that \( G_1 \) clearly separates ‘N’ from ‘T’ in the test set without any error. Thus, it is obvious that the 0% test error is not dependent upon the particular subset of the current test set. All these results strongly indicate that the Hotelling’s \( T^2 \) statistic, in particular, and the multivariate analysis, in general, is a valuable tool for the detection of DE genes in the microarray analysis.

4.4 ROC approach to the mixed dataset

Let \( D \) denote the differential expression between the tumor and the normal tissues, which is defined as \( D = Y - X \). Let \( D_0 \) denote the hypothetical version of \( D \) under the null hypothesis. Once distributions of \( D \) and \( D_0 \) are determined following the procedure described in the Appendix section, one can proceed calculating \( \text{ROC}_{\text{pair}}(q_0) \), with a suitably chosen \( c_0 \) value.

![Image of Figure 2](https://academic.oup.com/bioinformatics/article-abstract/21/4/517/203305/fig2)

**Table 2.** The number of genes (gene pairs) which yield 0% test error

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Based on the top 50 genes of ( t )-statistic</th>
<th>25 pairs of ( T^2 )-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLDA</td>
<td>( &gt;10 )</td>
<td>2</td>
</tr>
<tr>
<td>DQDA</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>SVML</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>SVMQ</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>SVMG</td>
<td>4</td>
<td>( &gt;10 )</td>
</tr>
</tbody>
</table>

DLDA and DQDA stand for diagonal linear discriminant analysis and diagonal quadratic discriminant analysis, respectively. SVML, SVMQ and SVMG denote support vector machines with linear, quadratic and Gaussian kernels, respectively.
### Table

<table>
<thead>
<tr>
<th>Gene Notation</th>
<th>Gene ID</th>
<th>Gene Name</th>
<th>Univariate t</th>
<th>Hotelling’s $T^2$</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$</td>
<td>AA634308</td>
<td>ATP-binding cassette, sub-family A (ABC1), member 8</td>
<td>-18.492</td>
<td>1372.29</td>
<td>0.749</td>
</tr>
<tr>
<td>$G_1$</td>
<td>AA487193</td>
<td>secreted frizzled-related protein 4</td>
<td>7.392</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_2$</td>
<td>AI655379</td>
<td>transient receptor potential cation channel, subfamily C, member 3</td>
<td>-20.889</td>
<td>1316.21</td>
<td>0.612</td>
</tr>
<tr>
<td>$G_2$</td>
<td>AI360919</td>
<td>a disintegrin and metalloproteinase domain 12 (meltrin alpha)</td>
<td>10.695</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 3.** Scatter plots of the top two gene pairs selected from the training set by Hotelling’s $T^2$ statistic are drawn based on the test set. $x$- and $y$-axes represent the log-intensity ratios, and ‘N’ and ‘T’ denote normal and tumor tissues, respectively. Details of the top two gene pairs are provided in the bottom table.
Fig. 4. Gene expression distributions for normal versus tumor specimens for the top ranked gene and the fourth ranked gene in terms of univariate $t$-statistic. We note that the fourth ranked gene has better separated the normal specimens from the tumor specimens than the top ranked gene.

Table 3. Top 20 genes in terms of ROC mix (1/36) values and their corresponding ranks in terms of $t_3$-statistic

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene id</th>
<th>Gene name</th>
<th>ROC mix (1/36)</th>
<th>Rank in terms of $t_3$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AA634308</td>
<td>ATP-binding cassette, subfamily A (ABC1), member 8</td>
<td>1.0000</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>AA452278</td>
<td>Solute carrier family 4, sodium bicarbonate co-transporter, member 4</td>
<td>1.0000</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>A665537</td>
<td>Transient receptor potential cation channel, subfamily C, member 3</td>
<td>0.9833</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>W46577</td>
<td>Endothelial cell-specific molecule 1</td>
<td>0.9833</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>A655374</td>
<td>Chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)</td>
<td>0.9833</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>AA046527</td>
<td>Transmembrane 4 superfamily member tetraspan NET-5</td>
<td>0.9750</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>AA010400</td>
<td>ets variant gene 4 (E1A enhancer binding protein, E1AF)</td>
<td>0.9750</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>AA425217</td>
<td>Cadherin 3, type 1, P-cadherin (placental)</td>
<td>0.9708</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>AA976699</td>
<td>Chromogranin A (parathyroid secretory protein 1)</td>
<td>0.9708</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>N93505</td>
<td>Transmembrane 4 superfamily member 2</td>
<td>0.9708</td>
<td>28</td>
</tr>
<tr>
<td>11</td>
<td>H23187</td>
<td>Carbonic anhydrase II</td>
<td>0.9708</td>
<td>40</td>
</tr>
<tr>
<td>12</td>
<td>W76339</td>
<td>Nuclear factor (erythroid-derived 2)-like 3</td>
<td>0.9625</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>A954935</td>
<td>Matrix metalloproteinase 11 (stromelysin 3)</td>
<td>0.9625</td>
<td>98</td>
</tr>
<tr>
<td>14</td>
<td>A925826</td>
<td>Inhibin, beta A (activin A, activin AB alpha polypeptide)</td>
<td>0.9583</td>
<td>31</td>
</tr>
<tr>
<td>15</td>
<td>R51912</td>
<td>Somatostatin</td>
<td>0.9542</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>AA704407</td>
<td>Extracellular link domain containing 1</td>
<td>0.9542</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>A1373245</td>
<td>Hypothetical protein FLJ21511</td>
<td>0.9542</td>
<td>50</td>
</tr>
<tr>
<td>18</td>
<td>W63749</td>
<td>B-cell CLL/lymphoma 2</td>
<td>0.9500</td>
<td>36</td>
</tr>
<tr>
<td>19</td>
<td>R93176</td>
<td>Carbonic anhydrase 1</td>
<td>0.9500</td>
<td>77</td>
</tr>
<tr>
<td>20</td>
<td>AA031514</td>
<td>Matrix metalloproteinase 7 (matrilysin, uterine)</td>
<td>0.9458</td>
<td>13</td>
</tr>
</tbody>
</table>

which, in turn, determines the false positive probability $t_0$ in the baseline distribution. In general $c_0$ is chosen to make $t_0$ very small. However, as Pepe et al. (2003) indicate, with a small number of tissue specimens, that the estimation of ROC($t_0$) at very small $t_0$ is not possible and hence in the real application one needs to compromise with the choice of $t_0$ such that it is small, but large enough to make ROC($t_0$) reasonably precise. Our choice of $t_0$ is 1/36. The top 20 genes in terms of ROC mix (1/36) values have 50% overlap with top 20 genes in terms of $t$-statistic. We noted that the top gene in terms of the $t$-statistic was ranked ninth in terms of ROC mix (1/36).

Table 3 shows the list of top 20 genes in terms of ROC mix (1/36) of Equation (3) and their corresponding ranks in terms of $t_3$-statistic of Equation (2).

4.5 Validation of the microarray gene expression

We selected six genes based on their ranks in $t$, $t_3$ and ROC mix values, and performed the RT–PCR experiment. The
list of genes, their ranks in terms of \( t \), \( t_3 \) and \( \text{ROC}_{\text{mix}} \), Pearson correlation between \( \log(\text{RT–PCR}) \) and microarray gene expression measurement, and \( \text{diff}_{\text{err}} \) of Equation (4) are given in Table 4. From Table 4, we note the small value of \( \text{diff}_{\text{err}} \) suggests that these two assays be interchangeably used for the prediction of a new tissue to be normal or tumor.

The second gene in Table 4 with the gene id AA634308 [ATP-binding cassette, subfamily A (ABC1), member 8] is the one associated with the 0% test error, which is the same gene denoted by \( G_1 \) in Figure 3. We note from Table 4 and Figure 3 that it has consistently high ranks in terms of four statistics we considered in this study, namely, \( t \), \( t_3 \), Hotelling’s \( T^2 \) statistic and \( \text{ROC}_{\text{mix}} \). Also in the earlier analysis of a subset of 58 cases (20 paired, 16 normal only and 22 tumor only cases), this same gene ranked the second, the top, and again the top, respectively, in the \( t \), \( t_3 \), and Hotelling’s \( T^2 \) statistic (Kim et al., 2005). The RT–PCR result of this gene is shown in Figure 1.

### 4.6 Subtype analyses

Each class of subtypes was not well separated as with normal versus tumor tissues. We used two sample \( t \)-statistic (with unequal variances) and employed FWER, FDR and \( p \)-FDR for controlling the Type I error due to multiple tests. We could detect nine DE genes between stages B and D with \( p \text{-FDR} = 0.33 \), and 36 DE genes for \( \text{CEA} \geq 5 \) versus \( \text{CEA} < 5 \) with \( p \text{-FDR} = 0.31 \). We failed to detect DE genes for colon cancer versus rectum cancer and for stages B versus C. These results are shown in Table 5.

The prediction rate for each class of subtypes based on the top 10 genes in terms of the \( t \)-statistic using SVM based on 10 pairs of training and test sets does not exceed 0.72, which corresponds to the rate for stages B versus D. We observed 0.49, 0.51, 0.66 and 0.64, for the prediction rates of colon cancer versus rectum cancer, stages B versus C, stages C versus D and \( \text{CEA} \geq 5 \) versus \( \text{CEA} < 5 \), respectively. However, so far as \( \text{CEA} \) is concerned we could improve the prediction rate by 17% through the SVM/regression-based gene selection approach described in Section 3. The prediction rates of regression approach alone and SVM/regression based approach were 0.64 and 0.75, respectively.

In the regression-based gene selection for predicting \( \text{CEA} \geq 5 \), we observed that the stepwise regression ended up either four gene or five gene model of Equation (6) in the Appendix section for each training set. Thus, we had 10 sets of finally selected genes, each set consisting of four or five genes. We counted the frequency of each gene belonging to 10 sets. Three genes have the frequency of 7 or more. These are AI088704\1|AI939310 (expressed sequence tags, ESTs), AA911045 (ESTs) and H68509 [alcohol dehydrogenase 6 (class V)].

### 5 DISCUSSION

We have developed statistical methods that are applicable to the mixed dataset of microarray experiment performed on human cancers by extending standard statistical procedures, such as \( t \)-statistic and ROC approach. The mixed dataset occurs quite often in clinical practice when the tissue material is not large enough to yield the adequate amount of RNA for undergoing the DNA microarray experiment.

Diverse cancer-related genes were included in the selected genes such as commonly up-regulated nuclear factor
erythroid 2-like 3 (NFE2L3), wingless-type MMTV integration site family (WNT5A), matrix metalloproteinase 7 (MMP 7), matrix metalloproteinase 11 (MMP 11), ets variant gene 4 (ETV4), and commonly down-regulated genes of chemokine (C-X-C motif) ligand 12 (CXCL 12), chromogranin A (CgA) and carbonic anhydrase II (CA II). High level of MMP 11 is associated with human cancer progression (Basset et al., 1997), and MMP 11 increased tumorigenesis through decreased cancer cell death with the help of apoptosis and necrosis (Boulay et al., 2001). In addition, MMP 7 is known to be overexpressed in human colorectal carcinomas (Ougolkov et al., 2002), and ETV4 can activate the promoters of various MMPs (Horiuchi et al., 2003). CA II and CA XII are the members of zinc metalloenzymes family and the loss of CA II expression is known to accompany the progression of colorectal carcinomas. (Indinnimeo et al., 2002). These information support the biological concept that the selected known genes and ESTs are related to colorectal cancer.

There has been a criticism on using univariate approaches for the detection of DE genes, since these approaches disregard the multidimensional structure of microarray data (Szabo et al., 2003). As an initial attempt of employing a multivariate analysis in the microarray data, we consider a random vector consisting of two genes and computed Hotelling’s T^2 statistic for all possible combinations of two gene pairs for the detection of DE genes between normal and tumor tissues. This multivariate approach provides a prediction rate that is at least as good as univariate approaches including the t-test. It was more sensitive than the univariate t-statistic for the detection of the gene that alone discriminated between normal and tumor tissues with 0% test error.

However, there are several issues that need to be addressed before Hotelling’s T^2 statistic is formally applied in the analysis of microarray data. The first issue would be calculating the P-value of the observed T^2 statistic. The second is to extend Hotelling’s T^2 statistic to a vector of length k ≥ 3. These two issues might be handled with random search algorithm suggested by Szabo et al. (2003) with the reasonable equipment of hardwares.

We reported that among the top 50 genes selected by the t-statistic, 44% were detected by Hotelling’s T^2 statistic. This 44% for the overlap proportion is in parallel with Szabo et al. (2003), who noted 41～47% overlap between gene sets detected by a multivariate method and several univariate approaches in an inhomogeneous dataset. Different gene sets detected by several univariate procedures and Hotelling’s T^2 statistic need to be further validated using Gene Ontology and molecular pathway findings.

In contrast to detecting an overwhelming number of DE genes between normal and tumor tissues, we failed to detect significant number of DE genes in subtype analyses. We detected nine DE genes between stages B and D and another 36 genes between CEA ≥5 and CEA <5 as provided in Table 5. However, the small sample size of stage D, high values of FDR and low-prediction rates all indicate that the current sample size, 68, is not large enough to detect the multiple mechanisms that underlie each class of subtypes.

The 87 colorectal cancer patients comprises three subgroups of sizes 36, 19 and 32. We did not find any statistical evidence against independence of these three subgroups with respect to age, gender, stage and CEA ≥5 versus CEA <5. Only for location (colon versus rectum) we obtained an unadjusted P-value of 0.017, which might yield Bonferroni adjusted P-value of 0.085. There is no clinical meaning at this point on this marginal significance of location versus three subgroups.

The presence of correlation between genes plays its role in Hotelling’s T^2 statistic for the selection of DE genes. As we discussed earlier we observed large values of Hotelling’s T^2 statistic mostly when either |t_1| or |t_2| was large, |t_1t_2| < 0, and |t_1t_2| was large in Equation (1). Therefore, one might expect that about half of the genes in the top 25 pairs of Hotelling’s T^2 statistic could be selected by the univariate t-statistic alone and the other half would not. This may raise the issue of inherent inconsistency between these two procedures. However, our data show that the top gene pair selected by Hotelling’s T^2 statistic contains the gene with the smallest test error.

We noted in this study that different gene sets provided more or less the same prediction rate. This aspect may be viewed from the multiplicity of model and integrating these different models would serve in future research.

ACKNOWLEDGEMENTS

B.S.K. was supported by a grant from the Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (02-PJ-12-G3-10411-00-03). S.H.L. was supported by a grant R04-203-000-10145-0 from the Basic Research Program of the Korea Science and Engineering Foundation. H.C.C. and S.Y.R. were supported by the Korea Science and Engineering Foundation (KOSEF) through the Cancer Metastasis Research Center (CMRC) at Yonsei University College of Medicine.

REFERENCES


**6 APPENDIX**

**6.1 Null distribution of \( D = Y - X \)**

Let \( D = Y - X \) and \( D_0 \) denote the hypothetical version of \( D \) under the null hypothesis of no differential expression. Let \( D \) denote the sample mean of \( D \) values based on \( n_1 \) observations. The distribution of \( D \) with a mean \( \delta_0 \) and the variance \( \sigma^2_0 \) is denoted by \( D \sim (\delta_0, \sigma^2_0) \). The distribution of \( D_0 \) is represented by \( D_0 \sim (\delta_0, \sigma^2_0) \). We augment the \( D \) notation by adding a superscript \( (i) \) to represent the \( i \)-th gene. Hence \( D^{(i)} \) denote \( D \) for the \( i \)-th gene and similarly for \( D^{(i)} \). We omit this superscript when the argument is based on each gene.

Let \( |[D^{(i)}]|_{\xi=1} \leq |D|_{\xi=1} \leq |D|_{\xi=p} \) denote the order statistics of \( |[D^{(i)}]|_{\xi=1} \leq |D|_{\xi=1} \leq |D|_{\xi=p} \). Where \( p \) is the number of genes spotted in a cDNA microarray. We assume, for simplicity, that \( D_0 \) has the same distribution with \( D \) except for the mean and variance. We expect that \( \delta_0 \leq \delta_0 \) and we do not necessarily assume that \( \delta_0 = 0 \) to allow a small baseline value that may vary depending on the experimental condition. We further assume that \( \sigma^2_0 \leq \sigma^2_0 \). There are several ways of estimating the distribution of \( D_0 \) using the matched pair sample data. We choose a set of genes, denoted by \( \mathcal{N}_p = \{ i \mid |[D^{(i)}]|_{\xi=1} < \varepsilon \} \) for a small \( \varepsilon > 0 \). The suitable choice of \( \varepsilon \) can be determined from the plot of...
\{m, \text{Var}(\overline{D}_{\lfloor m \rfloor})\}_{m=1}^p\}. We concluded from the plot (data not shown) that the first 100 order statistics provide information on the non-DE genes. We observed that beyond the 100th smallest $|D|$ value, the variance tended to increase slowly. Based on these 100 genes, we could estimate the mean and the variance of the null distribution.

6.2 Regression-based gene selection procedure for predicting CEA $\geq 5$

Define CEA$_j$ to be the CEA level for the $j$-th patient and let $M_{ij}$ denote the $i$-th gene expression level for the $j$-th patient for $j = 1, \ldots, 68$ and $i = 1, \ldots, p (p = 12,850)$. We describe the procedure below.

(1) For each $i$, we estimate the regression equation from the training set

$$\frac{1}{\sqrt{\text{CEA}_j}} = \beta_0 + \beta_1 I\{\text{Stage}_j = B\} + \beta_2 I\{\text{Stage}_j = C\} + \gamma M_{ij} + \varepsilon_j,$$

where $\text{Stage}_j$ represents the stage of the $j$-th patient, and $I\{A\}$ is an indicator function that assumes 1 if $A$ is true, and 0 otherwise. We found that the inverse of square root transformation of CEA$_j$ was most appropriate after applying Box–Cox transformations to a selected subset of genes.

(2) We found 27 significant genes with pFDR $<$ 0.31.

(3) Employing a stepwise regression using 27 genes in (2) resulted in a five gene model with $R^2 = 0.71.$

$$\frac{1}{\sqrt{\text{CEA}_j}} = \beta_0 + \beta_1 I\{\text{Stage}_j = B\} + \beta_2 I\{\text{Stage}_j = C\} + \sum_{i=1}^5 \gamma_i M_{ij} + \varepsilon_j.$$

(6)

The ranks of these five genes in terms of the $t$-statistic are 58, 63, 65, 1813 and 5029.

(4) Prediction rates of $\{\text{CEA} \geq 5\}$ were calculated using SVM and the regression Equation (6).

(5) We further randomly divided 68 tumors into a training set of 45 tumor and a test set of 23. For this new pair of training and test sets, we repeat the steps (1)–(4) with the same pFDR level, obtained the regression Equation (6) but with possibly different number of genes. Finally, we predicted the CEA levels for the test set using SVM and the estimated regression equation.

We iterated this procedure of Steps 1–5 for 10 times and averaged prediction rates.