Gene expression

Hotelling’s \( T^2 \) multivariate profiling for detecting differential expression in microarrays

Yan Lu\(^1\), Peng-Yuan Liu\(^1\), Peng Xiao\(^1\) and Hong-Wen Deng\(^2,1,3,^*\)

\(^1\)Osteoporosis Research Center, Creighton University, 601 N. 30th Street, Suite 6787, Omaha, NE 68131, USA,
\(^2\)The Key Laboratory of Biomedical Information Engineering of Ministry of Education and Institute of Molecular Genetics, School of Life Science and Technology, Xi’an Jiaotong University, Xi’an 710049, PR China and
\(^3\)Laboratory of Molecular and Statistical Genetics, College of Life Sciences, Hunan Normal University, Changsha, Hunan 410081, PR China

Received on October 13, 2004; revised and accepted on April 15, 2005
Advance Access publication May 19, 2005

ABSTRACT

Summary: The most widely used statistical methods for finding differentially expressed genes (DEGs) are essentially univariate. In this study, we present a new \( T^2 \) statistic for analyzing microarray data. We implemented our method using a multiple forward search (MFS) algorithm that is designed for selecting a subset of feature vectors in high-dimensional microarray datasets. The proposed \( T^2 \) statistic is a corollary to that originally developed for multivariate analyses and possesses two prominent statistical properties. First, our method takes into account multidimensional structure of microarray data. The utilization of the information hidden in gene interactions allows for finding genes whose differential expressions are not marginally detectable in univariate testing methods. Second, the statistic has a close relationship to discriminant analyses for classification of gene expression patterns. Our search algorithm sequentially maximizes gene expression difference/distance between two groups of genes. Including such a set of DEGs into initial feature variables may increase the power of classification rules. We validated our method by using a spike-in HGU95 dataset from Affymetrix. The utility of the new method was demonstrated by application to the analyses of gene expression patterns in human liver cancers and breast cancers. Extensive bioinformatics analyses and cross-validation of DEGs identified in the application datasets showed the significant advantages of our new algorithm.

Availability: The program for the method proposed in this article and supplementary material are available at [http://orclinux.creighton.edu/hotelling/index.htm](http://orclinux.creighton.edu/hotelling/index.htm)

Contact: deng@creighton.edu

Supplementary information: [http://orclinux.creighton.edu/hotelling/index.htm](http://orclinux.creighton.edu/hotelling/index.htm)

1 INTRODUCTION

Microarray technology is a powerful approach for genomic research, which allows the monitoring of expression profiles for tens of thousands of genes in parallel and is already producing huge amounts of data. Microarray data contain valuable information about gene functions, inter-gene dependencies and underlying biological processes, and open a new avenue for discovering gene co-regulations, gene interactions, metabolic pathways and gene–environment interactions, etc (Slonim, 2002). Microarray data are characterized with high dimensions and small sample sizes. Statistical inference from such high dimensional data structures is challenging (Mehta et al., 2004). Several data-mining methodologies, such as clustering analysis and classification techniques, have been widely used to analyze gene expression data for identifying groups of genes sharing similar expression profiles (Alon et al., 1999; Brazma and Vilo, 2000; Eisen et al., 1998; Golub et al., 1999).

While clustering and classification techniques have proven to be useful to search similar gene expression patterns, these techniques do not answer the most fundamental question in microarray experiments: which genes are responsible for biological differences between different cell types and/or states of a cell cycle? This question amounts to statistically testing the null hypothesis that there is no difference in expression under comparison. Various statistical methods have been proposed for identifying differentially expressed genes (DEGs), from using a simple fold change to using various linear models as well as Bayesian methods (Baldi and Long, 2001; Chen et al., 1997; Kerr et al., 2000; Tusher et al., 2001; Wang and Ethier, 2004; Wittenhall and Smyth, 2004); but none of them has yet gained widespread acceptance for microarray data analyses. A common characteristic of these methods is essentially of univariate nature.

It is well known that genes never act alone in a biological system—they work in a cascade of networks (Leung and Cavalieri, 2003). Expression profiles of multiple genes are often correlated and thus are more suitably modeled as mutually dependent variables in development of a statistical testing framework. However, most of the current statistical methods ignore the multidimensional structure of the expression data and fail to efficiently utilize the valuable information for gene interactions. Multivariate analyses take advantage of the correlation information and analyze the data from multiple genes jointly. As a result, multivariate statistical techniques are receiving increased attention in expression data analyses (Chilingaryan et al., 2002; Szabo et al., 2002). However, applications of well-established multivariate statistical techniques for microarray data analyses are not straightforward because of the unusual features of the microarray data, such as high dimensions and small sample sizes.
In this study, we presented a Hotelling’s $T^2$ test that utilizes multiple gene expression information to identify DEGs in two test groups. The Hotelling’s $T^2$ statistic is a natural multidimensional extension of the $t$-statistic that is currently a widespread approach for detecting DEGs in testing individual genes. The Hotelling’s $T^2$ method has been applied to various aspects of life sciences including genome association studies (Xiong et al., 2002), microarray process control (Model et al., 2002) and data control charts (Mason et al., 1995). We implemented our method using a multiple forward search (MFS) algorithm that is designed for selecting a subset of feature vectors in high-dimensional microarray datasets. A resampling-based technique was also developed to smooth statistical fluctuation in $T^2$ statistic owing to large variability inherent in microarray data and to accommodate experiments with missing values for various spots on microarrays. The proposed Hotelling’s $T^2$ method was validated by using a spike-in HGU95 dataset from Affymetrix. To illustrate its utility, we applied our method to the microarray data analyses of cancer genes in human liver cancers (Chen et al., 2002) and breast cancers (van’t Veer et al., 2002).

2 METHODS

2.1 Hotelling’s $T^2$ statistic

We consider a microarray experiment composed of $n_D$ samples from a disease group and $n_N$ samples from a normal group. Suppose that the expression levels of $J$ genes are measured and used as variables to construct a $T^2$ statistic. Let $X_i^D$ be the expression level for gene $j$ of sample $i$ from the disease group and $X_i^N$ be the expression level for gene $j$ of sample $k$ from the normal group. The expression level vectors for samples $i$ and $k$ from the disease and normal groups can be expressed as $X_i^D = (X_{i1}^D, \ldots, X_{iJ}^D)^T$ and $X_k^N = (X_{k1}^N, \ldots, X_{kJ}^N)^T$, respectively. The mean expression level vectors of $J$ genes in the disease and normal groups can be expressed as

$$\bar{X}_D^j = \frac{1}{n_D} \sum_{i=1}^{n_D} X_i^D \quad \text{and} \quad \bar{X}_N^j = \frac{1}{n_N} \sum_{k=1}^{n_N} X_k^N,$$

respectively. The mean expression level vectors for $J$ genes in the disease and normal groups can be expressed as $\bar{X}_D = (\bar{X}_1, \ldots, \bar{X}_J)^T$ and $\bar{X}_N = (\bar{X}_1, \ldots, \bar{X}_J)^T$, respectively. The pooled variance–covariance matrix of expression levels for $J$ genes for the disease and normal samples is then defined as

$$S = \frac{(n_D - 1)S_D + (n_N - 1)S_N}{n_D + n_N - 2},$$

where

$$S_D = \frac{1}{n_D} \sum_{i=1}^{n_D} (X_i^D - \bar{X}_D)(X_i^D - \bar{X}_D)^T,$$

and

$$S_N = \frac{1}{n_N} \sum_{k=1}^{n_N} (X_k^N - \bar{X}_N)(X_k^N - \bar{X}_N)^T.$$  \hspace{1cm} (1)

Let $S_D$ and $S_N$ be the variance–covariance matrix of expression levels for $J$ genes in the disease and normal groups, respectively. The covariance terms in $S_D$ and $S_N$ account for the correlation and interdependence (interactions) of gene expression levels.

Hotelling’s $T^2$ statistic for gene differential expression studies is then defined as

$$T^2 = \frac{n_D n_N}{n_D + n_N} (\bar{X}_D - \bar{X}_N) S^{-1} (\bar{X}_D - \bar{X}_N)^T.$$ \hspace{1cm} (2)

This statistic combines information from the mean and dispersion of all the variables (genes being tested) in microarray experiments. When we compare two groups of samples, both of which have a large sample size, see under the null hypothesis that the distributions in both groups are the same, i.e. there is no differential expression for any genes being tested in the disease and normal groups, the central limit theorem dictates that,

$$\frac{n_D + n_N - J - 1}{T^2} \sim F_{n_D + n_N - J - 1}$$

is asymptotically $F$-distributed with $J$ degrees of freedom for the numerator and $n_D + n_N - J - 1$ for the denominator.

2.2 Multiple forward search (MFS) algorithm

There are usually a relatively small number of independent samples used in microarray experiments, while a relatively large number of genes under comparison may actually be differentially expressed. Hence, the pooled sample variance–covariance matrix $S$ in $T^2$ statistic may be singular and not invertible. As a remedy for this problem, we propose a MFS algorithm that sequentially maximizes expression differences between groups of genes. This algorithm allows for iteratively and exhaustively finding a set of target DEGs. The basic structure of our MFS algorithm is outlined below:

Step 1. Calculate $T^2$ statistics for each of all the genes that are measured in datasets, and find the gene $j_1$ that maximizes $T^2$, denoting $T^2_{j_1}$.

Step 2. If $P$-value$_{T^2_{j_1}} < \alpha$ (a predefined significance level), calculate $T^2$ statistics for two groups, one is the gene $j_1$, and the other is one of the remaining genes excepting the gene $j_1$. Find the gene $j_2$ that maximizes $T^2$ combining with the gene $j_1$, denoting $T^2_{j_1,j_2}$.

Step 3. If $P$-value$_{T^2_{j_1,j_2}} < P$-value$_{T^2_{j_1}}$, repeat Step 2 by adding one more gene that maximizes $T^2$ combining with the genes $j_1$ and $j_2$.

Step 4. Repeat Step 3 until $P$-value$_{T^2_{j_1,j_2,\ldots,j_n}} > P$-value$_{T^2_{j_1,j_2,\ldots,j_{n-1}}}$ or the number of genes is larger than $n_1 + n_2 - 2$. Then the selected genes $j_1, j_2, \ldots, j_{n-1}$ are the first subset of identified DEGs.

Step 5. Exclude gene $j_1, \ldots, j_{n-1}$, and repeat Steps 1–4.

Step 6. Repeat Step 5 until the $P$-value of $T^2$ statistic of the starting gene is larger than $\alpha$, i.e. $P$-value$_{T^2_{1,2,\ldots,J}} > \alpha$, and stop searching.

The structure of our algorithm that adds one gene to $T^2$ statistics in each of the Steps 1–4 allows for a fast updated computation of $T^2$ statistics by avoiding matrix inverse. For example, the pooled variance–covariance matrix for $n$ genes can be partitioned into a block form:

$$S_n = \begin{bmatrix} S_{n-1} & a^T \\ a & b \end{bmatrix},$$

where $S_{n-1}$ is the variance–covariance matrix for $n - 1$ genes, $a$ is the covariance vector for the first $n - 1$ genes and the $n$-th gene, $a^T$ is the transpose of $a$, and $b$ is the variance for the $n$-th gene. The inverse of $S_n$ is then calculated as

$$S_n^{-1} = \begin{bmatrix} S_{n-1}^{-1} - \frac{1}{k}a^TS_{n-1}^{-1}a & -\frac{1}{k}S_{n-1}^{-1} \\ -\frac{1}{k}S_{n-1}^{-1}a^T & \frac{1}{k} \end{bmatrix},$$

where $k = b - a^TS_{n-1}^{-1}a$. We only need to invert $S_1$ for the starting gene and recursively calculate the inverse of $S_n$ by the above formula (4). Singular value decomposition can be used to get the general inverse of covariance matrix when the iterative method fails.

2.3 Resampling

In microarray experiments, it is not unusual that some data points are missing due to poor quality. Large inherent ‘noise’ in microarray data also renders estimation of the pooled sample variance–covariance matrix unrobust and
variable. We propose the following procedure to handle such incomplete multivariate data that are used in constructing Hotelling’s $T^2$ statistics used in the MFS algorithm:

Step 1. Resample $N$ replicates with replacement of subjects from the original data, denoting a replicate as $R_r$, $r = 1, \ldots, N$.

Step 2. Calculate $T^2$ statistic for each $R_r$, denoting $T^2_r$, $r = 1, \ldots, N$.

Step 3. Exclude 5% the lower and upper tails of $T^2_r$, and obtain the mean $T^2$ using the remaining $T^2_r$.

The number of replicates $N$ could be as large as possible but is limited by computational capability, however. This resampling strategy can alleviate statistical fluctuation and reduce sampling errors. It can help identify a consistent subset of DEGs.

3 VALIDATION OF HOTELLING’S $T^2$ STATISTIC

3.1 Human genome U95 spike-in dataset

In the course of developing and validating the Affymetrix Microarray Suite (MAS) 5.0 algorithm, Affymetrix produced and provided data (containing 12 640 genes) from a set of 59 arrays (HGU95) organized in a Latin-square design (http://www.affymetrix.com/support/technical/sample_data/datasets.affx). This dataset consists of 14 spike-in gene groups at known concentrations in 14 experimental groups, comprising 12 groups of three replicates (A–L) and two groups of 12 replicates (group M–P and group Q–T). In our analyses, the latter two groups of 12 replicates were used to validate our method for identifying DEGs. The correlations among 14 genes, measured by their concentrations, ranged from $-0.965$ to $0.924$. Since a single RNA source was used, any probe-set not in the list of 14 genes should be negative for differential expression. Conversely, all of the probes in the list of 14 genes should be positive for differential expression.

3.2 Identification of DEGs

The probe levels were obtained by Robust Multi-array Analysis (RMA) (Irizarry et al., 2003) and MAS 5.0, respectively. The significance level 0.001 was adopted in the analyses. The $t$-test resulted in 4 false negatives and 7 false positives using the expression levels from RMA, and 2 false negatives and 33 false positives using the expression levels from MAS 5.0. The Hotelling’s $T^2$ method resulted in 1 false negative and 6 false positives using the expression level from RMA, and 2 false negatives and 16 false positives using the expression levels from MAS 5.0 (Table 1).

4 APPLICATION EXAMPLES

4.1 Human liver cancers

4.1.1 Dataset

We applied our method to publicly available datasets from the study of Chen et al. (2002) who examined gene expression patterns in human liver cancers by cDNA microarrays containing 23 075 clones representing $\sim 17 400$ genes. In their study, they profiled genomic expressions in $>200$ samples including 102 primary hepatocellular carcinoma (HCC) from 82 patients and 74 non-tumor liver tissues from 72 patients. Two-sample Welch $t$-statistic was adopted to identify DEGs between two sets of samples. Permutation procedure was used to determine $P$-values. Genes with permutation $P$-values $<0.001$ were considered to be differentially expressed. Data are available online at the Stanford Microarray Database (SMD; http://genome-www5.stanford.edu/).
We chose genes for our analyses using the same standard as the original study (Chen et al., 2002). Namely, all non-flagged array elements for which the fluorescent intensity in each channel was > 1.5 times the local background were considered well measured. Genes for which fewer than 75% of measurements across all the samples met this standard were excluded from further analyses (http://MedGENE.org, http://www.genscript.com/cgi-bin/products/genome.cgi). In the following, primary data processing, we performed the following three comparisons using Hotelling’s $T^2$-test:

1. **HCC versus non-tumor liver tissues.** We compared gene expression in 82 HCC versus 74 non-tumor liver tissue samples. 11,386 genes with good measured data for more than 62 HCC and more than 56 normal livers were used in our analyses.

2. **HCC with negative versus positive p53 staining.** To investigate the relationship between p53 mutations and gene expression programs in HCC, Chen et al. (2002) examined 59 HCC specimens by immunohistochemical staining for p53 protein. They found 23 of these HCC analyzed have positive p53 staining, which has been noted to correlate with p53 mutation or inflammation in HCC (Hsu et al., 1993). 11,744 genes with good measured data for more than 27 HCC with negative p53 staining and more than 27 HCC with positive p53 staining were used in our analyses.

3. **HCC with versus without venous invasion.** To identify the role of vascular invasion in tumor spread and metastasis, 81 tumor samples were classified by histopathological evaluation as 43 negative and 38 positive for vascular invasion (Chen et al., 2002). 11,161 genes with good measured data for more than 32 HCC without venous invasion and more than 29 HCC with venous invasion were used in our analyses.

### 4.1.2 Identification of DEGs

Table 2 summarizes the total numbers of DEGs identified in the above three sub-datasets by the two-sample Welch $t$ statistic in the original paper (Chen et al., 2002) and those by the Hotelling’s $T^2$-statistic we proposed. The Hotelling’s $T^2$ statistic found more DEGs than the $t$-test. In the comparison of the HCC versus non-tumor tissues, more than 2000 DEGs were identified by both of the two methods. In the comparison of the HCC with positive versus negative p53 staining, two-thirds of DEGs declared by the $t$-test were also identified by our Hotelling’s $T^2$ statistic. However, a relatively smaller proportion of the DEGs is shared by the two methods in the comparison of the HCC with versus without venous invasion.

Nine bioinformatics resources were used to evaluate the relative performance of the two statistical methods on the basis of our known knowledge about the gene function that has been obtained by various aspects of empirical studies, including biochemical, genetic, epidemiological, pharmacological and physiological. These nine bioinformatics resources include KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.ad.jp/kegg/kegg2.html), MedGENE database (http://ipu.peq.med.harvard.edu/MEGENE/), OMIM (Online Mendelian Inheritance in Man™, http://www.ncbi.nlm.nih.gov/Omim/), Atlas of Genetics and Cytogenetics in Oncology and Haematology (http://www.infobiogen.fr/services/chromcancer/index_genes.gc.html), CancerGene Database (http://carroll.vjf.cnrs.fr/cancergene/), Cancer GeneWeb (http://www.cancertindex.org/geneweb/X070601.htm), MTB (Mouse Tumor Biology Database, http://tumor.informatics.jax.org/FMPro), GeneCards™ (http://bioinfo.weizmann.ac.il/cards/) and GenePool (http://www.genscript.com/cgi-bin/products/genome.cgi). In the following, we detail our bioinformatics analyses for the DEGs identified by each of the two methods in the comparison of the HCC versus non-tumor tissues.

Table 2. Comparison of the DEGs discovered by the two methods in human liver cancers

<table>
<thead>
<tr>
<th>Pathways</th>
<th>Only $t$-test</th>
<th>Only $T^2$-test</th>
<th>Both methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC versus non-tumor (82/74)</td>
<td>3964</td>
<td>4508</td>
<td>2051</td>
</tr>
<tr>
<td>HCC with negative versus positive p53 staining (36/23)</td>
<td>121</td>
<td>146</td>
<td>83</td>
</tr>
<tr>
<td>HCC without versus with venous invasion (43/38)</td>
<td>91</td>
<td>151</td>
<td>34</td>
</tr>
</tbody>
</table>

Table 3 shows grouping categories of the DEGs according to their relationships with HCC as cited in literature using the database MedGENE. We classified the DEGs identified by the two methods into several different categories (Hu et al., 2003b): (1) first-degree

<table>
<thead>
<tr>
<th>Pathways</th>
<th>Only $t$-test</th>
<th>Only $T^2$-test</th>
<th>Both methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>p53 signaling pathway</td>
<td>8</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Wnt signaling pathway</td>
<td>21</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>16</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>Apoptosis profile</td>
<td>12</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Angiogenesis</td>
<td>15</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>MAP kinase signaling pathway</td>
<td>24</td>
<td>39</td>
<td>57</td>
</tr>
<tr>
<td>Growth factors</td>
<td>7</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Signal transduction in cancer</td>
<td>20</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>DNA damage signaling pathway</td>
<td>13</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>Stress and toxicity</td>
<td>17</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>Metastasis</td>
<td>16</td>
<td>13</td>
<td>22</td>
</tr>
<tr>
<td>DNA repair pathway</td>
<td>4</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>TGFβ BMP signaling pathway</td>
<td>14</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>JAK/STAT signaling pathway</td>
<td>13</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>231</td>
<td>309</td>
</tr>
</tbody>
</table>

Note: In Tables 3 and 4, ‘both methods’ indicates the number of DEGs discovered by both the Hotelling’s $T^2$ and $t$-tests. ‘Only $t$-test’ indicates the number of DEGs that were only found by $t$-test but not by the Hotelling’s $T^2$-test. While ‘Only Hotelling’s $T^2$’ indicates the number of DEGs that were only found by the Hotelling’s $T^2$ test but not by $t$-test.
associations, i.e. genes that have been directly linked to this disease by gene term search; (2) first-degree associations by gene family term, i.e. genes that have been directly linked to this disease by gene family term search; and (3) second-degree associations, i.e. genes that have never been co-cited with this disease but have been linked (in the same pathway) to at least one first-degree gene. Others are genes that have not been previously associated with this disease. The two-sample Welch t-test identified 26 and 42 DEGs more than the Hotelling’s $T^2$ test in the categories of first-degree associations and first-degree associations by gene family term, respectively. However, the Hotelling’s $T^2$ test identified 166 more DEGs than the t-test in the category of second degree associations. These results indicate that the Hotelling’s $T^2$ test is sensitive to finding genes whose differential expression is not detectable marginally, in addition to the genes found by the one-dimensional criteria. The names of these DEGs are detailed in Supplementary Table 2. In Supplementary Table 2, the statistic for ranking the gene list is relative risk of disease (Hu et al., 2003b).

In our analyses, several reported novel and important cancer-associated genes that were differentially expressed between the HCC and normal tissues were identified by the Hotelling’s $T^2$ test but not by the t-test. These genes included, for example, pleiomorphic adenoma gene-like 2 (PLAGL2), budding uninhibited by benzimidazoles 1 homolog beta (BUB1B) and budding uninhibited by benzimidazoles 3 homolog (BUB3), centromere protein F (CENPF), hepatoma-derived growth factor (HDGF), interleukin 1 beta (IL1B) and catenin (cadherin-associated protein) beta 1 (CTNNB1). PLAGL2 is a zinc-finger protein that recognizes DNA and/or RNA. It displays typical biomarkers of neoplastic transformation: (1) lose cell–cell contact inhibition, (2) show anchorage-independent growth and (3) induce tumors in nude mice (Hansen et al., 2002). BUB1B and BUB3 are the mitotic checkpoint genes, which were found over expressed in gastric cancer and associated with tumor cell proliferation (Cahill et al., 1998; Grabsch et al., 2003). CENPF was identified as antigen-inducing novel antibody responses during the course of transition to malignancy (Casio et al., 1993). RT–PCR and western blot analyses detected increased HDGF expression in malignant hepatoma cell lines (Hu et al., 2003a). Polymorphisms in the IL-1B-511 genetic locus are one of the possible determinants of progression of hepatitis C to HCC (Tanaka et al., 2003). The expression of CTNNB1 is highly correlated with tumor progression and postoperative survival in HCC (Inagawa et al., 2002). A recent study also reported a new non-canonical pathway through which Wnt-5a antagonizes the canonical Wnt pathway by promoting the degradation of beta-catenin (Topol et al., 2003).

### 4.2 Human breast cancers

#### 4.2.1 Dataset

We applied our method to another dataset from the study of van’t Veer et al. (2002). This dataset contains totally 78 primary breast cancers: 34 from patients who developed distant metastases within 5 years and 44 from patients who continued to be disease-free after a period of at least 5 years. These data were collected for the purpose of finding a prognostic signature of breast cancers in their gene expression profiles.

---

**Table 4.** Categories of the DEGs discovered by the two methods for the comparison of HCC versus non-tumor tissues

<table>
<thead>
<tr>
<th>Categories</th>
<th>Only t-test</th>
<th>Only $T^2$-test</th>
<th>Both methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-degree associations</td>
<td>165</td>
<td>139</td>
<td>318</td>
</tr>
<tr>
<td>First-degree associations by gene family term</td>
<td>127</td>
<td>85</td>
<td>156</td>
</tr>
<tr>
<td>Second-degree associations</td>
<td>509</td>
<td>675</td>
<td>571</td>
</tr>
<tr>
<td>Total</td>
<td>801</td>
<td>899</td>
<td>1045</td>
</tr>
</tbody>
</table>

**Table 5.** Error rates of discriminant analyses with DEGs discovered by the two methods in human liver tissues

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Number of genes used</th>
<th>Error rates t-test</th>
<th>$T^2$-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC versus non-tumor (82/74)</td>
<td>63</td>
<td>0.51</td>
<td>0.07</td>
</tr>
<tr>
<td>HCC with negative versus positive p53 staining (36/23)</td>
<td>42</td>
<td>0.55</td>
<td>0.16</td>
</tr>
<tr>
<td>HCC without versus with venous invasion (43/38)</td>
<td>56</td>
<td>0.51</td>
<td>0.23</td>
</tr>
</tbody>
</table>

---

and the cross validation was repeated 10 times. Each time, one of the 10 subsets was used as a validation set and the other 9 subsets were put together to form a learning set. We applied the most significant genes identified in the learning sets by each of the two methods to classify samples in the validation sets using the following discriminant analyses. The average error rate across all 10 trials was then computed.

The discriminant analyses were based on Mahalanobis distance that accounts for ranges of acceptability (variance) between variables and compensates for interactions (covariance) between variables (Taguchi and Jugulum, 2002). The distance between two samples can be calculated as, $D^2(T_1, G_1) = (X - \mu_i)^T V_i^{-1} (X - \mu_i)$ where $i = 1, 2$ represent group 1 and 2, respectively. $D^2$ is the generalized squared distance of the test sample $T$ from the i-th group $G_i$. $V_i$ is the within-group covariance matrix of the i-th group. $\mu_i$ is the vector of means of gene expression levels of the i-th group. $X$ is the vector of gene expression levels observed in the test sample $T$. If $D^2(T_1, G_1) < D^2(T_2, G_2)$, the test sample is from group 1, otherwise from group 2.

The average error rates for classification of samples for the two statistics are listed in Table 5. The average error rates from the Hotelling’s $T^2$ test are always lower than those from the t-test in all of the three sub-datasets when the numbers of genes used for classification are the same. Furthermore, the average error rates of the Hotelling’s $T^2$ test were much more stable than the t-test when different numbers of DEGs are used for classification. For example, for the HCC versus non-tumor tissues, the average error rates of the Hotelling’s $T^2$ test varied from 0.05 to 0.07 when the numbers of genes used for classification varied from 29 to 63, while the average error rates of the t-test varied from 0.17 to 0.51. When using the $T^2$ test, the average error rates for the sub-datasets of the HCC with negative versus positive p53 staining and the HCC with versus without venous invasion were higher than that for the HCC versus non-tumor tissues. This is due to the fact that the sample sizes of learning sets in the former two datasets are much smaller than the latter one.
4.2.2 Cluster analysis van’t Veer et al. (2002) found 70 optimal marker genes as ‘prognosis classifier’. This classifier predicted correctly the outcome of disease for 65 out of the 78 patients (83%), with five poor prognosis and eight good prognosis patients, respectively, assigned to the opposite category (see the Figure 2b in the study of van’t Veer). We used our Hotelling’s $T^2$ test to find the signature genes and applied the top 70 and 52 genes to classify the respective samples. Hierarchical clustering was used for classifying these 78 primary breast cancers. We obtained 8 out of 78 incorrect classifications, with 4 poor prognosis and 4 good prognosis patients assigned to the opposite category when using the 70-genes classifier (Fig. 1a); while we obtained 5 out of 78 incorrect classifications with 1 poor prognosis and 4 good prognosis patients assigned to the opposite category, when using the 52-gene classifier (Fig. 1b).

5 DISCUSSION

The exponential growth of gene expression data is accompanied by an urgent need for theoretical and algorithmic advances in integrating, analyzing and processing the large amount of valuable information. The Hotelling’s $T^2$ statistic presented in this paper is a novel tool for analyzing microarray data. The proposed $T^2$ statistic is a corollary to its original counterpart developed for multivariate analyses (Hotelling, 1947). Our statistic for microarray data analysis possesses two prominent statistical properties. First, our method takes into account multidimensional structure of microarray data. The utilization of the information for gene interactions allows for finding genes whose differential expressions are not marginally detectable in univariate testing methods. Second, the statistic has a close relationship to discriminant analyses for classification of gene expression patterns. Our search algorithm sequentially maximizes gene expression difference/distance between two comparison groups. Inclusion of such a set of DEGs into initial feature variables may increase the power of classification rules. In addition, the Hotelling’s $T^2$ test gives one $P$-value for groups of genes rather than a $P$-value for each gene. Multiple testing problems are hence much less serious than univariate testing methods, unless too many groups of genes are tested.

We first validated our method by using a spike-in HGU95 dataset from Affymetrix. This dataset consists of 14 spike-in genes at known concentrations. The Hotelling’s $T^2$ method gave fewer false positives and negatives than t-test. For example, our method produced 1 false negative and 6 false positives, while the t-test produced 4 false negatives and 7 false positives, using the expression levels from RMA. We then applied our method to the analyses of gene expression patterns in human liver cancers (Chen et al., 2002). Extensive bioinformatic analyses and cross-validation of the DEGs identified in the study illustrated several significant advantages over the univariate t-test. First, our method discovered more DEGs in pathways related to HCC. Though $P$-values of the t-statistics for some genes do not exceed the threshold for significance level in our study, these genes contribute significantly to the Hotelling’s $T^2$ statistic. These genes per se usually show marginal differential expressions but are correlated with other strong DEGs (Chilingaryan et al., 2002). Second, our method identified significantly more DEGs in the category of second-degree associations with HCC. Interestingly, these genes have never been co-cited with HCC in the literature but have been linked to at least one gene that has been directly linked to this disease by gene term search, using the MedGene database (Hu et al., 2003b). Their roles in HCC tumors are worthy of further examination. This may also reflect a potential bias in the literature, where either HCC was investigated on an individual gene-by-gene study or multiple-gene expression profiles were analyzed univariately. In this sense, it is not surprising that the t-test found more DEGs in the categories of first-degree associations since current available databases seem to unduly favor the univariate approaches such as the t-test here. Third, we found several novel cancer-associated genes such as PLAGL2 and BUB1B that were also highly expressed in HCC tumors. They play important roles in the process of tumor formation and development as evidenced in a considerable body of literature (Cahill et al., 1998; Casiano et al., 1993; Grabsch et al., 2003; Hu et al., 2003a; Inagawa et al., 2002; Tanaka et al., 2003; Topol et al., 2003). Fourth, we reduced the misclassification to as low as 5% using $\leq 30$ genes identified by the $T^2$ tests. This holds great promise in clinical diagnoses and classification of tumors. A large number of training samples, such as those we have examined here, are imperative to establish highly predictive classification functions. Finally, we applied our method to find signature genes in human breast cancers (van’t Veer et al., 2002). These signature genes used in hierarchical clustering resulted in higher accuracy of predicting disease outcome.

Recently, several authors proposed multivariate approaches for selecting subsets of DEGs (Chilingaryan et al., 2002; Szabo et al., 2002). In their pioneering efforts, a measure of distance between vectors of gene expressions is defined for simultaneously comparing a set of genes. However, the final chosen DEGs are dependent on the choice of the predetermined cluster size, $k$; larger values of $k$ typically lead to a larger list of DEGs found in a study (Szabo et al., 2002); while in our method, a variable entering into $T^2$ statistic depends on its excess contribution to the maximization of two-group differences, thus circumventing some of the problems inherent in variable selection. This was implemented by a heuristic MFS algorithm that is designed for selecting a subset of feature vectors in high-dimensional microarray datasets. Another prominent feature of our method is that we developed a resampling-based approach to smooth out statistical fluctuations of the $T^2$ statistic owing to large variability inherent in microarray data which is very helpful for finding a consistent set of DEGs. The resampling technique can also be used for handling incomplete multivariate data, a common problem in microarray experiments.

More recently, Goeman et al. (2004) developed a score test for testing whether or not some pre-specified groups of genes are differentially expressed. The groups of genes could be those that are involved in a particular biochemical pathway or a genomic region of interest and should be specified before testing. This method is very valuable for testing some known pathways that affect clinical outcome in combination with groups of genes. However, the score statistic merely tests gene-expression differences in pre-specified groups of genes between two tissue types and is not intended for group wise search for DEGs. Intuitively, the Hotelling’s $T^2$ statistic can also be used for testing pre-specified group differences and its relative performance requires further investigations in comparison with the score statistic.

One of the stopping rules associated with our search algorithm is the number of genes entering into the $T^2$ statistic is smaller than $n_1 + n_2 - 2$, where $n_1$ and $n_2$ are sample sizes of two different tissue types under comparison. This restriction can be released by using principle component analysis (PCA) to reduce the dimensionality of variables entered in the Hotelling’s $T^2$ statistic (Mardia et al.,...
Fig. 1. Hierarchical clustering analyses of 78 primary human breast tumors using the 70-gene classifier (a) and the 52-gene classifier (b). Each row represents a single gene and each column represents a single tumor. <5 represents patients developing distant metastases within 5 years; >5 represents patients continuing to be disease-free after a period of at least 5 years. The samples marked with dashes were incorrectly classified.
Fig. 1. Continued.
1979). Within the framework of stabilized multivariate tests (Lauter et al., 1996), the Hotelling’s $T^2$ statistic can be performed on the basis of linear scores that are derived from the original variables using PCA. Incorporating the PCA technique into our method is valuable, especially for microarray experiments with small sample sizes. However, a small sample size may not warrant multivariate normal distribution of the data, in which permutation tests using the $T^2$ statistic may be necessary.

**ACKNOWLEDGEMENTS**

Investigators of this work were partially supported by grants from NIH, State of Nebraska and CNSF. The study was also benefited by grants from Huo Ying Dong Education Foundation, the Ministry of Education of China, Xi’an Jiaotong University and Hunan Normal University.

**REFERENCES**


