Selecting Genes by Test Statistics

Dechang Chen,1 Zhenqiu Liu,2 Xiaobin Ma,3 and Dong Hua4

1Division of Epidemiology and Biostatistics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA
2Bioinformatics Cell, TATRC, 110 North Market Street, Frederick, MD 21703, USA
3Department of Computer Science and Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, USA
4Department of Computer Science, The George Washington University, 801 22nd St. NW, Washington, DC 20052, USA

Received 28 April 2004; revised 22 November 2004; accepted 23 November 2004

Gene selection is an important issue in analyzing multiclass microarray data. Among many proposed selection methods, the traditional ANOVA F test statistic has been employed to identify informative genes for both class prediction (classification) and discovery problems. However, the F test statistic assumes an equal variance. This assumption may not be realistic for gene expression data. This paper explores other alternative test statistics which can handle heterogeneity of the variances. We study five such test statistics, which include Brown-Forsythe test statistic and Welch test statistic. Their performance is evaluated and compared with that of F statistic over different classification methods applied to publicly available microarray datasets.

INTRODUCTION

Microarrays provide information about the expression level of the genes represented on the array. Such gene expression profiling has been successfully applied to class prediction, where the purpose is to classify and predict the diagnostic category of a sample by its gene expression profile [1, 2, 3, 4]. Various machine learning methods are currently used for class prediction. However, the task of prediction by microarrays is challenging, due to a large number of genes (features) and a small number of samples involved in the problem. As a consequence, one has to identify a small subset of informative genes contributing most to the classification task. Performing feature selection is essential for microarray prediction problems, since high-dimensional problems usually involve higher computational complexity and bigger prediction errors.

Many methods have been proposed to select informative genes. One category of such work depends on the traditional t test statistic [5, 6, 7] and analysis of variance (ANOVA) F test statistic [8, 9]. While t is used for two-class prediction problems, F is used for multiclass problems. The statistics t and F are not only used in class prediction, they also apply to the class discovery [10, 11]. The main goal of class discovery is to identify subtypes of diseases. The major difference between class prediction and class discovery is that the former uses labeled samples while the latter uses unlabeled samples.

Although t and F have been commonly used in the analysis of gene expression data, there exists a misunderstanding on the roles of t and F. The test statistic t is used to detect the difference between the means of two populations and it has two versions depending on whether or not the two variances of the two populations are equal. The test statistic F is often used to detect the difference among the means of three or more populations under the assumption that the variances of the involved populations are equal. Of course, the F statistic can be used to detect the difference between the means of two populations. In doing this, one can show that the F statistic is equivalent to the t statistic based on the equal variance, that is, one procedure rejects the null hypothesis that the two populations have the same mean if and only if the other procedure rejects the null hypothesis. In analyzing gene expression data, the t statistic is based on unequal variances so that its extension will never reach the ANOVA F. Therefore, for multiclass prediction problems, it is natural to explore other statistics which do not assume equal variances.

© 2005 Hindawi Publishing Corporation
In this paper, we study the effect on multiclass prediction results of gene selection from six test statistics: ANOVA F test statistic, Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, Cochran test statistic, and Kruskal-Wallis test statistic. The five last test statistics can be viewed as extensions of the t statistic used in two-class prediction problems. Their performance will be compared with that of the F statistic.

This paper is organized as follows. In “models and methods,” we describe the statistical model for gene expression values and describe test statistics on the classification results by using our selection approach and different machine learning techniques, applied to five publicly available microarray datasets. Our conclusion is given in “conclusion.”

MODELS AND METHODS

In this section, we will first introduce a general statistical model for gene expression values and describe test statistics for testing the equality of the class means. We then present our approach to select genes using power and correlation.

Statistical model

Assume there are \( k \) (\( \geq 2 \)) distinct tumor tissue classes for the problem under consideration and there are \( p \) genes (inputs) and \( n \) tumor mRNA samples (observations). Suppose \( X_{gs} \) is the measurement of the expression level of gene \( g \) from sample \( s \) for \( g = 1, \ldots, p \) and \( s = 1, \ldots, n \). In terms of an expression matrix \( G \), we may write

\[
G = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{p1} & X_{p2} & \cdots & X_{pn}
\end{pmatrix}.
\]

It is seen that the columns and rows of the expression matrix \( G \) correspond to samples and genes, respectively. Note that \( G \) is a matrix consisting of data highly processed through preprocessing techniques that include image analysis and normalization and often logarithmic transformations. We assume that the data \( G \) are standardized so that the genes have mean 0 and variance 1 across samples. Given a fixed gene, let \( Y_{ij} \) be the expression level from the \( j \)th sample of the \( i \)th class. Note that these \( Y_{ij} \) come from the corresponding row of \( G \). For example, for gene 1, \( Y_{ij} \) are a rearrangement of the first row of \( G \). We consider the following general model for \( Y_{ij} \):

\[
Y_{ij} = \mu_i + \epsilon_{ij}, \quad \text{for } i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, n_i
\]

with \( n_1 + n_2 + \cdots + n_k = n \). In the model, \( \mu_i \) is a parameter representing the mean expression level of the gene in class \( i \), \( \epsilon_{ij} \) are the error terms such that \( \epsilon_{ij} \) are independent random normal variables, and

\[
E(\epsilon_{ij}) = 0, \quad V(\epsilon_{ij}) = \sigma_i^2 < \infty,
\]

for \( i = 1, \ldots, k; \quad j = 1, 2, \ldots, n_i \). Schematically, the expression levels \( Y_{ij} \) look like the following:

| Classes | \( Y_{11} \) | \( Y_{21} \) | \( Y_{31} \) | \cdots | \( Y_{k1} \) |
|--------|-------------|-------------|-------------|-------|-------------|
| \( Y_{12} \) | \( Y_{22} \) | \( Y_{32} \) | \cdots | \( Y_{k2} \) |
| \vdots | \vdots | \vdots | \ddots | \vdots |
| \cdots | \cdots | \cdots | \cdots | \cdots |
| \( Y_{1n_1} \) | \cdots | \cdots | \cdots | \( Y_{kn_k} \) |

Note that if the variances are equal, that is, \( \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \), then the above model is simply the commonly used one-way ANOVA model. For the microarray data, we believe that heterogeneity in the variances is more realistic, since different \( \sigma_i \) may describe different variations of the gene expression across classes.

One of the main tasks associated with the above model is to detect whether or not there is some difference among the means \( \mu_1, \mu_2, \ldots, \mu_k \). For the case of homogeneity of variances, the well-known ANOVA F test is the optimal test to accomplish the task [12, 13]. However, with heterogeneity of the variances, the task is challenging and is closely related to the well-known Behrens-Fisher problem [14]. When the sample sizes in all classes are equal, that is, \( n_1 = n_2 = \cdots = n_k \), the presence of heterogeneous variances of the errors only slightly affects the F test. When the sample sizes are unequal, the effect is serious [15]. The actual type-I error is inflated if smaller sizes \( n_i \) are associated with larger variances \( \sigma_i^2 \). In addition, the significance levels are smaller than anticipated if larger sizes \( n_i \) are associated with larger variances \( \sigma_i^2 \). The above indicates that for our model, the F test may not be appropriate for testing \( H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \) versus \( H_1 \): not all the \( \mu_i \) are equal. Therefore some alternatives to the F test are worthy of investigating.

Test statistics

After introducing the statistical model for gene expression values, we now turn to the test statistics used to test the equality of the class means for a fixed gene. We will consider the following six test statistics. The first five are parametric test statistics, while the last one is nonparametric.

(a) ANOVA F test statistic. The definition of this test is

\[
F = \frac{(n-k) \sum_{i=1}^{n_i} (Y_{i.} - \bar{Y}_{..})^2}{(k-1) \sum_{i=1}^{k} (n_i - 1) \sigma_i^2},
\]

where \( \bar{Y}_{i.} = \sum_{j=1}^{n_i} Y_{ij} / n_i \), \( \bar{Y}_{..} = \sum_{i=1}^{k} n_i \bar{Y}_{i.} / n \), and \( \sigma_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 / (n_i - 1) \). For simplicity, we use \( \sum \) to indicate the sum is taken over the index \( i \). Under \( H_0 \) and assuming variance homogeneity, this well-known test statistic has a distribution of \( F_{k-1,n-k} \) [13].
(b) Brown-Forsythe test statistic [16]. This is given by

\[ B = \frac{\sum n_i (\bar{Y}_i - \bar{Y}_.)^2}{\sum (1 - n_i/n) s_i^2}. \]  

Under \( H_0 \), \( B \) is distributed approximately as \( F_{k-1, w} \), where

\[ y = \frac{\left[ \sum (1 - n_i/n) s_i^2 \right]^2}{\sum (1 - n_i/n)^2 s_i^4/(n_i - 1)}. \]  

(c) Welch test statistic [17]. This is defined as

\[ W = \frac{\sum w_i (\bar{Y}_i - \bar{Y}_.)^2}{(k-1)+2(k-2)(k+1)^{-1} \sum (n_i - 1)^{-1} (1-h_i)^2}, \]  

with \( w_i = n_i/s_i^2 \) and \( h_i = w_i/\sum w_i \). Under \( H_0 \), \( W \) has an approximate distribution of \( F_{k-1,w} \), where

\[ v_w = \frac{k^2 - 1}{3 \sum (n_i - 1)^{-1} (1-h_i)^2}. \]  

(d) Adjusted Welch test statistic [18]. It is similar to the Welch test statistic and defined to be

\[ W^* = \frac{\sum w_i^* (\bar{Y}_i - \bar{Y}_.)^2}{(k-1)+2(k-2)(k+1)^{-1} \sum (n_i - 1)^{-1} (1-h_i^*)^2}, \]  

where \( w_i^* = n_i/(\phi_i s_i^2) \) with \( \phi_i \) chosen such that \( 1 \leq \phi_i \leq (n_i - 1)/(n_i - 3) \), and \( h_i^* = w_i^*/\sum w_i^* \). Under \( H_0 \), \( W^* \) has an approximate distribution of \( F_{k-1,v_W} \), where

\[ v_w^* = \frac{k^2 - 1}{3 \sum (n_i - 1)^{-1} (1-h_i^*)^2}. \]  

In this paper, we choose \( \phi_i = (n_i + 2)/(n_i + 1) \), since this choice provides reliable results for small sample sizes \( n_i \) and a large number \( k \) of populations [18].

(e) Cochran test statistic [19]. This test statistic is simply the quantity appearing in the numerator of the Welch test statistic \( W \), that is,

\[ C = \sum w_i (\bar{Y}_i - \bar{Y}_.)^2, \]  

where \( w_i \) and \( h_i \) are given in (c). Under \( H_0 \), \( C \) has an approximate distribution of \( \chi^2_{k-1} \).

(f) Kruskal-Wallis test statistic. This is the well-known nonparametric test and is given by

\[ H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1), \]  

where \( R_i \) is the rank sum for the \( i \)th class. The ranks assigned to \( Y_{ij} \) are those obtained from ranking the entire set of \( Y_{ij} \) (use the average rank in case of tied values). Assuming each \( n_i \geq 5 \), then under \( H_0 \), \( H \) has an approximate distribution of \( \chi^2_{k-1} \) [20].

**Gene selection**

With the test statistics introduced above, we are able to discuss the issue of gene selection. It has been well demonstrated in the literature that gene selection is an important issue in microarray data analysis. It is also known that with a large number of genes (usually in thousands) present, no practical method is available to locate the best set of genes, that is, the smallest subset of genes that offer optimal prediction accuracy. In this paper, the focus lies in comparing the performance of different test statistics in selecting genes for the classification of tumors based on gene expression profiles. Identifying a gene selection process to achieve good classification results is not the purpose of this paper. To make the comparison straightforward, we adopt the simplest gene selection approach as follows. First, we formulate the expression levels of a given gene by a one-way ANOVA model, as shown in “statistical model.” We then use the test statistics in “test statistics” to determine the power of genes in discriminating between tumor types. Given a test statistic \( \mathcal{F} \), we define the discrimination power of a gene as the value of \( \mathcal{F} \) evaluated at the \( n \) expression levels of the gene. This definition is based on the fact that with larger \( \mathcal{F} \) the null hypothesis \( H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \) will be more likely rejected. Therefore, the higher the discrimination power is, the more powerful the gene is in discriminating between tumor types. Finally, we choose as informative genes those genes having high power of discrimination.

We note that the discrimination power of genes could be determined equally well by the \( p \) value from \( \mathcal{F} \). However, due to small sizes \( n_i \), it is hard to justify the approximation of the known distribution to \( \mathcal{F} \). Therefore the \( p \) values may not reflect the actual functionality of \( \mathcal{F} \). This drawback is overcome by using the value of \( \mathcal{F} \) to determine the power of discrimination. Another obvious benefit is that using the value of \( \mathcal{F} \) will greatly simplify the calculation.

In [18], extensive simulations have been conducted to examine the behavior of some test statistics for testing the equality of population means. The test statistics studied include \( B, W, W^*, F, \) and \( C \). The results show that with homogeneity of the variances, the ANOVA F test is the optimal test, as stated in “statistical model.” However, this assumption of homogeneity is rarely met in practice. Under heterogeneity of variances, the simulation results in [18] show that the test statistics \( B, W, \) and \( W^* \) provide acceptable control of type I errors. This implies that the genes identified by \( B, W, \) and \( W^* \) are more likely to be powerful than those by \( F \) and \( C \) in discriminating between tumor types, and thus the prediction errors resulting from \( B, W, \) and \( W^* \) are expected to be lower than those from \( F \) and \( C \). The nonparametric test statistic \( H \) can be applied to data with less restriction, for example, ordinal data, and thus is expected to perform worse than test statistics such as \( B, W, W^* \), and \( C \). The above discussion will be further verified by our experiments on gene expression data conducted in “experimental results.”
EXPERIMENTAL RESULTS

In this section we investigate the effect on gene selection of the six test statistics introduced in “test statistics.” Five gene expression datasets and five prediction methods are used for this purpose. The performances of the test statistics are evaluated in terms of class prediction errors.

Datasets

We considered five multiclass gene expression datasets: leukemia72 [1], ovarian [21], NCI [22, 23], lung cancer [24], and lymphoma [25]. Table 1 presents more details of the datasets.

Comparison of test statistics

The gene selection procedure described above depends on the test statistic. Given a gene selection process from a test statistic, different classification methods may lead to different prediction errors. In our experiments, we used the following five prediction methods: naive Bayes, nearest neighbor, linear perceptron, multilayer perceptron neural network with 5 nodes in the middle layer, and support vector machines with a second-order polynomial kernel. All the algorithms are from Matlab PRTools 3.01 by Robert P. W. Duin.

To calculate the overall prediction error, we used leave one out (LOO) cross-validation. For a dataset with \( n \) samples, this method involves \( n \) separate runs. For each of the runs, \( n - 1 \) data points are used to train the model and then prediction is performed on the remaining data point. The overall prediction error is the sum of the errors on all \( n \) runs.

Table 2 presents a comparison of the six test statistics when 50 informative genes were used. In the table, F, B, W, \( W^* \), C, and H represent the ANOVA F test statistic, Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, Cochran test statistic, and Kruskal-Wallis test statistic, respectively. The first number in each cell denotes the average of 5 prediction errors from 5 different classification methods. The second number in each cell is the median of the 5 prediction errors. The results in the table suggest that B, W, \( W^* \), and, C perform better than F and H. Similar to Table 2, Tables 3 and 4 present comparison results with 100 and 200 informative genes, respectively.

Results in Tables 2, 3, and 4 may be summarized in a way by figures. Consider the average errors in the tables. For a fixed dataset and fixed number of informative genes, the performances of the six test statistics can be ranked. The fifteen ranks achieved by a test statistic could be used to obtain a 95% confidence interval of the mean rank for the test statistic. The corresponding bar chart plotting six confidence intervals is given in Figure 1. The bar chart based on the median errors in Tables 2, 3, and 4 is presented in Figure 2. Clearly, both figures show that B, W, \( W^* \), and C outperform F and H. These results indicate that the proposed models in “statistical model” without assuming equal variances are preferred to those assuming equal variances.

We note that in the above experiments, the performance of C is comparable to those of B, W, and \( W^* \). This does not look consistent with the discussion in “gene selection.” One reason might be that we only examined 5 datasets in this paper. Our opinion is that if more data sets are explored, the overall performance of C should be worse than that of B, W, or \( W^* \). We leave this as our future work.

Before concluding, we point out that it is useful to assess the importance of genes selected by the test statistics from the biological perspective. Since this is not the focus of our research work in this paper, below we only provide a simple example to examine some genes selected by the Brown-Forsythe test statistic for the leukemia dataset. This dataset was also studied by Getz et al [26]. They extracted the stable clusters of genes by the coupled two-way clustering analysis and concluded that those genes grouped into the same cluster share certain biological significance such as on the same pathway. Among the top 50
### Table 1. Multiclass gene expression datasets.

| Dataset       | Leukemia72 | Ovarian | NCI   | Lung cancer | Lymphoma |
|---------------|------------|---------|-------|-------------|----------|
| No of genes   | 6817       | 7129    | 9703  | 918         | 4026     |
| No of samples | 72         | 39      | 60    | 73          | 96       |
| No of classes | 3          | 3       | 9     | 7           | 9        |

### Table 2. Performances of the test statistics with 50 informative genes.

| Dataset  | F   | B   | W   | W*  | C   | H   |
|----------|-----|-----|-----|-----|-----|-----|
| Leukemia | 3.4 | 2.4 | 2.8 | 2.8 | 3.2 | 3.0 |
|          | 3   | 2   | 3   | 3   | 3   | 3   |
| Ovarian  | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
|          | 0   | 0   | 0   | 0   | 0   | 0   |
| NCI      | 36.0| 32.0| 27.4| 26.0| 27.0| 35.4|
|          | 35  | 29  | 27  | 27  | 27  | 35  |
| Lung cancer | 17.6 | 17.0 | 17.6 | 17.6 | 18.0 | 18.0 |
|          | 17  | 17  | 18  | 18  | 18  | 18  |
| Lymphoma | 23.8| 19.8| 14.0| 14.0| 12.8| 22.0|
|          | 23  | 19  | 12  | 12  | 13  | 20  |

### Table 3. Performances of the test statistics with 100 informative genes.

| Dataset  | F   | B   | W   | W*  | C   | H   |
|----------|-----|-----|-----|-----|-----|-----|
| Leukemia | 3.4 | 3.0 | 3.0 | 3.0 | 3.2 | 3.0 |
|          | 3   | 3   | 4   | 3   | 3   | 3   |
| Ovarian  | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
|          | 0   | 0   | 0   | 0   | 0   | 0   |
| NCI      | 33.0| 22.6| 23.8| 25.2| 25.2| 31.6|
|          | 33  | 22  | 25  | 26  | 26  | 31  |
| Lung cancer | 12.2 | 12.2 | 11.4 | 12.2 | 12.2 | 15.8 |
|          | 12  | 12  | 11  | 11  | 11  | 14  |
| Lymphoma | 21.8| 19.2| 13.0| 13.8| 14.4| 18.2|
|          | 17  | 16  | 12  | 12  | 12  | 18  |

### Table 4. Performances of the test statistics with 200 informative genes.

| Dataset  | F   | B   | W   | W*  | C   | H   |
|----------|-----|-----|-----|-----|-----|-----|
| Leukemia | 3.0 | 3.0 | 2.4 | 2.8 | 1.8 | 2.4 |
|          | 3   | 3   | 2   | 3   | 1   | 2   |
| Ovarian  | 0.4 | 0.2 | 0.2 | 0.2 | 0.2 | 0.4 |
|          | 0   | 0   | 0   | 0   | 0   | 0   |
| NCI      | 25.6| 22.6| 22.6| 22.8| 22.2| 25.6|
|          | 26  | 22  | 24  | 25  | 24  | 25  |
| Lung cancer | 15.2 | 12.6 | 14.2 | 13.2 | 12.8 | 13.2 |
|          | 13  | 11  | 12  | 12  | 12  | 11  |
| Lymphoma | 21.2| 18.8| 12.0| 12.6| 12.8| 16.2|
|          | 15  | 14  | 8   | 9   | 8   | 14  |
Table 5. Mapping from genes selected by the Brown-Forsythe test statistic for the leukemia data to clusters of genes of interest provided by Getz et al [26].

| Gene description                                                                 | Access number | Cluster by Getz et al [26] | B          |
|---------------------------------------------------------------------------------|---------------|---------------------------|------------|
| GB DEF = T-cell antigen receptor gene T3-delta                                   | X03934        | LG5                       | 70.808014  |
| Protein tyrosine kinase related mRNA sequence                                    | L05148        | LG5                       | 43.676056  |
| CD33CD33 antigen (differentiation antigen)                                       | M23197        | LG1                       | 42.435883  |
| GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA| U23852 s      | LG5                       | 35.120228  |
| T-cell surface glycoprotein CD3 epsilon chain precursor                          | M23323 s      | LG5                       | 35.028965  |
| CTSD (cathepsin D) (lysosomal aspartyl protease)                                 | M63138        | LG1                       | 34.865067  |
| HLA class II histocompatibility antigen, DR alpha chain precursor                | X00274        | LG6                       | 31.882597  |
| HLA class I histocompatibility antigen, F alpha chain precursor                  | X17093        | LG6                       | 31.83585   |
| Leukotriene C4 synthase (LTC4S) gene                                            | U50136 rna1   | LG1                       | 31.183104  |
| RNS2 (ribonuclease 2) (eosinophil-derived neurotoxin (EDN))                     | X16546        | LG1                       | 29.52516   |
| TIMP2 (tissue inhibitor of metalloproteinase 2)                                  | M32304 s      | LG1                       | 28.233025  |
| LMP2 gene extracted from Homo sapiens genes TAP1, TAP2, LMP2, LMP7, and DOB     | X66401 cds1   | LG6                       | 27.11849   |

Informative genes from the Brown-Forsythe test statistic, 12 were mapped to the clusters of genes of interest given in [26]. Table 5 shows the information about the gene names, access numbers, corresponding clusters as well as the values of the Brown-Forsythe statistic. For details on the explanation of biological significance of clusters LG1, LG5, and LG6, readers are referred to [26].

CONCLUSION

In this paper, we have compared the performance of different test statistics in selecting genes for multi-classification of tumors using gene expression data. Experiments show (a) the model for gene expression values without assuming equal variances is more appropriate than that assuming equal variances; (b) Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, and Cochran test statistic perform much better than ANOVA F test statistic and Kruskal-Wallis test statistic.

DISCLAIMER

The opinions expressed herein are those of the authors and do not necessarily represent those of the Uniformed Services University of the Health Sciences and the Department of Defense.

ACKNOWLEDGMENTS

The authors thank Dr. Hanchuan Peng of Lawrence Berkeley National Laboratory for providing the NCI, lung cancer, and lymphoma data. The authors also thank the referees for providing many valuable comments. D. Chen was supported by the National Science Foundation grant CCR-0311252.

REFERENCES

[1] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–537.
[2] Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA. 2001;98(26):15149–15154.
[3] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Amer Statist Assoc. 2002;97(457):77–87.
[4] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunked centroids of gene expression. Proc Natl Acad Sci USA. 2002;99(10):6567–6572.
[5] Xiong M, Jin L, Li W, Boerwinkle E. Computational
methods for gene expression-based tumor classification. BioTechniques. 2000;29(6):1264–1270.

[6] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002;18(1):39–50.

[7] Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. Genome Inform Ser Workshop Genome Inform. 2002;13:51–60.

[8] Ghosh D. Singular value decomposition regression models for classification of tumors from microarray experiments. In: Proceedings of the 2002 Pacific Symposium on Biocomputing. Lihue, Hawaii: 2002:18–29.

[9] Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics. 2002;18(9):1216–1226.

[10] Ding C. Analysis of gene expression profiles: class discovery and leaf ordering. In: Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2002). Washington, DC: 2002:127–136.

[11] Li W, Fan M, Xiong M. SamCluster: An integrated scheme for automatic discovery of sample classes using gene expression profile. Bioinformatics. 2003;19(7):811–817.

[12] Lehman EL. Testing Statistical Hypotheses. 2nd ed. NY: Wiley; 1986.

[13] Neter J, Kutner MH, Nachtsheim CJ, et al. Applied Linear Statistical Models. 4th ed. Chicago, Ill: McGraw-Hill; 1996.

[14] Stuart A, Ord JK, Arnold S. Kendall's Advanced Theory of Statistics. Volume 2A: Classical Inference and the Linear Model. 6th ed. London: Edward Arnold; 1999.

[15] Montgomery DC. Design and Analysis of Experiments. 5th ed. NY: Wiley; 2001.

[16] Brown MB, Forsythe AB. The small sample behavior of some statistics which test the equality of several means. Technometrics. 1974;16:129–132.

[17] Welch BL. On the comparison of several mean values: An alternative approach. Biometrika. 1951;38:330–336.

[18] Hartung J, ArGA D, Makambi KH. Small sample properties of tests on homogeneity in one-way ANOVA and meta-analysis. Statist Papers. 2002;43:197–235.

[19] Cochran WG. Problems arising in the analysis of a series of similar experiments. J R Stat Soc Ser C Appl Stat. 1937;4:102–118.

[20] Daniel WW. Biostatistics: A Foundation for Analysis in the Health Sciences. 7th ed. NY: Wiley; 1999.

[21] Welsh JB, Zarrinkar PP, Sapinoso LM, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci USA. 2001;98(3):1176–1181.

[22] Ross DT, Scherf U, Eisen MB, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. 2000;24(3):227–235.

[23] Scherf U, Ross DT, Waltham M, et al. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000;24(3):236–244.

[24] Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci USA. 2001;98(24):13784–13789.

[25] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503–511.

[26] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci USA. 2000;97(22):12079–12084.