STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2069–2087
Published online 1 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2082

Gene selection for microarray data analysis using principal component analysis

Antai Wang*,† and Edmund A. Gehan

Department of Biomathematics and Biostatistics, Georgetown University, Lombardi Cancer Center, Washington, DC, U.S.A.

SUMMARY

Principal component analysis (PCA) has been widely used in multivariate data analysis to reduce the dimensionality of the data in order to simplify subsequent analysis and allow for summarization of the data in a parsimonious manner. It has become a useful tool in microarray data analysis. For a typical microarray data set, it is often difficult to compare the overall gene expression difference between observations from different groups or conduct the classification based on a very large number of genes. In this paper, we propose a gene selection method based on the strategy proposed by Krzanowski. We demonstrate the effectiveness of this procedure using a cancer gene expression data set and compare it with several other gene selection strategies. It turns out that the proposed method selects the best gene subset for preserving the original data structure. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: microarray data; principal component analysis; procrustes criterion; gene selection

1. INTRODUCTION

In a typical microarray data analysis, several groups of tissues may be available each of which is characterized by a large number of gene expressions and an objective is to derive a classification rule based upon the gene expressions. Usually the number of genes, p, is very large and this makes it extremely difficult to utilize all the gene expression values in a classification rule. Therefore, a crucial step in the microarray data analysis is first to select a subset of the most informative genes and then to derive a classification rule that discriminates among groups of observations and minimizes classifications errors.

Suppose now that there are two groups of samples. The most commonly used approach for selecting a subset of genes is to perform a two-sample t-test or Wilcoxon Mann–Whitney test for each gene and then rank the genes based on the observed test statistics or the corresponding p-values. Satagopan and Panageas [1] have described this approach in detail. The t-test

*Correspondence to: Antai Wang, Ph.D. Suite 183, Building D, 4000 Reservoir Rd, NW, Washington, DC 20057-1484, U.S.A.
†E-mail: aw94@georgetown.edu

Received May 2004
Accepted September 2004

based approach assumes that gene expression values of each gene are approximately normally distributed and independent (thus they do not take into account any correlation among genes). Owing to multiple testing, an adjustment is needed for the raw p-values. The genes with significant adjusted p-values (e.g. <0.01 or 0.05) could be identified as the most informative genes. According to this criterion, the most important genes are those that are most differentially expressed. One concern about this approach is that while the method may be able to identify genes that can distinguish samples from different groups, those selected genes may not be able to preserve the overall features of the original data set very well.

Besides the t-test based method, there are also some other strategies for performing the gene selection. Jolliffe [2] has proposed a procedure based on principal component analysis (PCA) that chooses a variable with the largest absolute coefficient from each of the leading principal components. This procedure has been found to be quite effective in selecting informative variables [3]. The idea of his strategy can be extended: several genes (variables) (instead of just one gene) with the largest absolute coefficients from each leading principal component can be retained in the subset of selected genes (see Reference [4]).

The purpose of this paper is to propose an alternative approach to perform the gene selection. The proposed method is also based on PCA and it is built upon a variable selection strategy suggested by Krzanowski [5]. As stated in K's paper: 'the best subset of variables will contain those variables which reproduce as closely as possible the general features of the complete data', so we apply his strategy to select genes that can also meet this requirement.

Suppose we have observed expression levels of p genes on each of n samples. PCA linearly transforms the p gene expression variables to corresponding principal component scores. Because the rank of a $n \times p$ matrix cannot exceed the minimum of n and p, we expect that the top M principal component scores (here $M < \min\{n, p\}$ and is defined as the essential dimensionality of the original data, see Section 2) will convey 'most of the sample information' inherent in the original p genes. Similarly, for any selected subset of genes (assuming that the number of genes in this subset is larger than M), we also expect that the top M principal component scores will convey 'most of the sample information' inherent in this subset of genes. Because our goal is to select a subset of genes which can preserve the overall features of the original data set, based on above discussion, this is equivalent to retaining a subset of genes whose top M principal component scores have the shortest 'distance' to the top M principal component scores of the original data. In Reference [5], the variable selection criterion developed to meet this requirement is also called 'Procrustes criterion'. The distance metric we use is the square of the Euclidean distance metric computed on the standardized expression profiles which is proportional to the one minus Pearson correlation distance metric.

In this paper, we will use the same backward elimination procedure introduced by Krzanowski to select genes. Because K's procedure involves the iterative calculations of the principal component scores and the data matrices of the microarray data are usually of high dimensionality, the method of obtaining principal component scores described in K's paper appears to be too computationally intensive to accomplish. To overcome this difficulty, we propose an alternative approach to simplify his procedure. It turns out that our way of calculating PCs is easy to implement and this makes K's method a useful tool to perform the gene selection for microarray data sets.

Our paper is organized in the following way: Section 2 contains a simple way to determine the essential dimensionality of a microarray data set. Our application of Krzanowski's method to perform the gene selection is then described in Section 3. A new stopping rule to determine

the optimum number of genes to select is presented in Section 4. Our proposed approach is demonstrated in Section 5 for a NCI cancer data set. We compare the proposed method to two alternative gene selection methods: the t-test based method and Jolliffe's method in Section 6 and end our paper with some remarks in Section 7.

2. THE DETERMINATION OF THE ESSENTIAL DIMENSIONALITY

Using the same notation as in Reference [5], suppose that the expression level of p genes $x_1,\ldots,x_p$ are observed on each of n samples ($n \ll p$), so that the resultant values can then be displayed in an $(n \times p)$ data matrix $X=(x_1,\ldots,x_p)$. We also assume that X has been mean centered and standardized. The essential dimensionality M is then defined as the number of principal components needed to be retained to account for the variation in the original data X. Therefore, before performing the gene selection, we need to determine the 'essential dimensionality' (M) of X. The sample correlation matrix can be expressed as $S=[1/(n-1)]X'X$. PCA provides p new orthogonal variables $z_1,\ldots,z_p$, where $z_i=\sum_{j=1}^{p} w_{ij}x_j$, $\mathrm{var}(z_i)=\lambda_i$ for $i=1,\ldots,p$ and $\mathrm{cov}(z_i,z_j)=0$ for $i\neq j$. Writing $Z=(z_1,\ldots,z_p)$, $w_i=(w_{i1},\ldots,w_{ip})'$, $W=(w_1,\ldots,w_p)$ and $L=\mathrm{diag}(\lambda_1,\lambda_2,\ldots,\lambda_p)$ where $\lambda_1\geqslant\cdots\geqslant\lambda_p\geqslant 0$, we can obtain the spectral decomposition of S:

$$S=WLW'$$

where W is an orthogonal matrix such that $WW'=W'W=I_p$ and principal component scores are given by the $(n\times p)$ matrix $Z=XW$. The singular value decomposition of X is

$$X=UDQ'$$

where U is an $n\times n$ matrix satisfying $U'U=UU'=I_n$, D is an $n\times n$ diagonal matrix with $d_{11}\geqslant\cdots\geqslant d_{nn}\geqslant 0$ and Q is an $n\times p$ matrix satisfying $QQ'=I_n$. This decomposition is unique apart from corresponding sign changes in U and Q. Therefore, if we write the (i,j)th elements of the X, U and Q as x_{ij}, u_{ij} and q_{ij}, respectively, the singular value decomposition has an element-wise representation

$$x_{ij}=\sum_{t=1}^{n} u_{it}d_t q_{jt}$$

Thus if the data structure is essentially M-dimensional ($M<\min\{n,p\}$), then the variation in the remaining $(n-M)$ principal components can be treated as random noise (see Reference [5]).

There are several strategies proposed to determine the essential dimensionality of a data set, see Reference [6]. In this paper, we have chosen to apply the most commonly used rule for choosing M. That is: to select a cumulative percentage of total variation which one desires that the selected PCs account for, say 80 or 90 per cent. The required number of PCs is then the smallest value of M for which this chosen percentage 80 or 90 per cent is equalled or exceeded.

3. GENE SELECTION PROCEDURE

Suppose that the essential dimensionality of a microarray data set has been determined as M in some way and we desire to find a procedure to select q genes ($M < q < p$) which can preserve the original data structure. As mentioned in Section 1, we wish to determine a gene subset whose first M principal component scores have the shortest distance to the first M principal component scores of the original microarray data set.

Denote X and $\tilde{X}$ ($n \times q$ matrix) as the original data set and the selected q-variable (gene) subset, respectively. Let $Y = (y_1, \ldots, y_M)$ and $\tilde{Z} = (z_1, \ldots, z_M)$ stand, respectively, for the $(n \times M)$ data matrix of top M principal component scores (corresponding to M largest eigenvalues of $S = [1/(n-1)]X'X$ and $\tilde{S} = [1/(n-1)]\tilde{X}'\tilde{X}$). $\tilde{Z}$ is therefore the M-dimensional approximation of q genes to the original data configuration. Let the singular value decomposition of $\tilde{Z}'Y$ be

$$\tilde{Z}'Y = U_1 \Sigma_1 V_1'$$

Sibson [7] has shown that for any two centroid-at-origin configurations Y and $\tilde{Z}$, the shortest squared distance between these two configurations under the transformations of Euclidean space generated by translation, rotation and reflexion is

$$\text{DIST}^2 = \text{Trace}\{YY' + \tilde{Z}\tilde{Z}' - 2\Sigma_1\}$$

The Procrustes criterion is then defined as the variable selection criterion by which one can select the 'best' subset of q variables which yields the smallest value of DIST^2 among all q-variable subsets [8, 5]. We use this criterion and follow the same backward elimination procedure introduced in K's paper to perform our gene selection:

1. Initially set $q = p$, and for fixed M, compute the matrix of principal component scores Y. Set $Z = Y$.
2. Obtain and store the matrix of principal component scores by deleting in turn each gene (variable) from Z.
3. Compute DIST^2 for each such matrix of scores and identify the gene x_u which yields the smallest DIST^2. Let $\tilde{Z}_{(u)}$ denote the corresponding matrix of scores.
4. Delete gene x_u. Set $Z = \tilde{Z}_{(u)}$ and return to step 2 with $p - 1$ genes. Continue this cycle until only q genes are left.

In Steps 1 and 2, we need to compute the principal component scores Y and $\tilde{Z}$ (corresponding to the M largest eigenvalues of X and $\tilde{X}$, respectively). We can retrieve them by performing either the spectral decomposition of $X'X$ and $\tilde{X}'\tilde{X}$ or the singular value decomposition of X and $\tilde{X}$. However, since p is very large, this means that we have to obtain the spectral decomposition or the singular value decomposition of matrices with dimensions close to $p \times p$ or $n \times p$ in the beginning of this procedure. Because we also need to do these calculations iteratively, Krzanowski's algorithm to compute the principal component scores via the singular value decomposition of X or $\tilde{X}$ directly appears to be too computationally intensive to realize.

Fortunately, we have found that the problem can be circumvented easily. Because $X'X$ and XX' have the same non-zero eigenvalues, $G = [1/(n-1)]XX'$ ($n \times n$ matrix) has the same non-zero eigenvalues as $S = [1/(n-1)]X'X$ ($p \times p$ matrix). Based on this fact, an alternative

way of obtaining principal component scores Y (an $n \times M$ matrix) is proposed as follows (the procedure of obtaining $\tilde{Z}$ based on $\tilde{X}$ is similar).

Denote λ_i as the ith largest non-zero eigenvalue for matrix G. Let v_i denote the standardized orthogonal eigenvector corresponding to eigenvalue λ_i. We then have

$$Gv_i = [1/(n-1)]XX'v_i = \lambda_i v_i \tag{1}$$

Left-multiplying both sides by X', we obtain

$$[1/(n-1)]X'X(X'v_i) = \lambda_i(X'v_i) \tag{2}$$

From above equation, it is easy to see that $X'v_i$ is an eigenvector of matrix S corresponding to the same eigenvalue λ_i. It follows from Equation (1) that the length of $X'v_i$ is simply $\sqrt{\lambda_i \times (n-1)}$. Let $h_i = (X'v_i)/\sqrt{\lambda_i \times (n-1)}$, it is also easy to check that h_i is an eigenvector corresponding to λ_i and the ith principal component scores of X can be simplified as

$$y_i = Xh_i = X(X'v_i)/\sqrt{\lambda_i \times (n-1)} = \sqrt{\lambda_i \times (n-1)}v_i \tag{3}$$

If we write the above results in a matrix form, we obtain

$$Y = V_1\Lambda$$

where

$$\Lambda = \sqrt{n-1} \times \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_M})$$

and

$$V_1 = (v_1, \ldots, v_M), \text{ here } M \leqslant \min\{n, p\}$$

For microarray data sets, the sample size n is usually much smaller than the number of genes p. Consequently, the matrix Y of top M principal component scores is much quicker to obtain (via V_1 and Λ) if we perform the spectral decomposition of XX' ($n \times n$ matrix) instead of doing the spectral decomposition of $X'X$ ($p \times p$ matrix) or singular value decomposition of X ($n \times p$ matrix). Therefore, applying this approach in steps 1 and 2 of the backward elimination procedure will greatly speed our gene selection process.

4. THE DETERMINATION OF THE OPTIMUM NUMBER OF SELECTED GENES

As pointed out by Krzanowski [9], if our selected gene subset contains only the top M genes, this gene subset may not be large enough to recover the original data structure very well. In practice, there is also a need to retain more than M genes for future study such as the class prediction based on the selected genes. We certainly do not need to retain all the genes because many of them give us redundant information. In fact, only a small proportion of all the genes are needed to preserve the data structure quite well and the variation explained by other genes can be treated as random noise.

Krzanowski [9] has proposed a stopping rule for the backward elimination process presented in Section 3. His stopping rule is used to determine the optimum number (denoted by $q^\star$) of variables to preserve data structure. He has also conducted a small Monte Carlo study to show

that the stopping rule chooses the right number of variables in general for data with special structures. However, his stopping rule is based on several model assumptions for the first M principal components of both the original and the selected data sets. One concern about it is that these assumptions are hard to verify. Another concern about this stopping rule is that it may be too conservative: applying this rule may result in keeping too many variables (genes) (as will be shown in Section 5).

Alternatively, we propose a new stopping rule based on the following criterion: the optimum number of selected genes can be determined such that adding one or more genes will not reduce the measure of the distance between the corresponding principal component scores (such as DIST^2) dramatically. The following procedure to determine this optimum number of selected genes can be programmed easily and has been shown to be quite effective:

1. Denote $X_{n\times q}$ as the selected subset of size q with minimum $\mathrm{DIST}^2=\mathrm{DIST}_q^2$ and $X_{n\times(q-1)}$ as the subset of size $q-1$ with minimum $\mathrm{DIST}^2=\mathrm{DIST}_{q-1}^2$ (It is also a subset of $X_{n\times q}$ from the backward selection procedure described in previous section). Define

$$R_q=\frac{(\mathrm{DIST}_{q-1}^2-\mathrm{DIST}_q^2)}{(\mathrm{DIST}_q^2)/(p-q)}$$

   for each $q\geqslant k$.
2. The optimum number $q^\star$ is chosen as the largest integer q satisfying $R_q>1.0$.

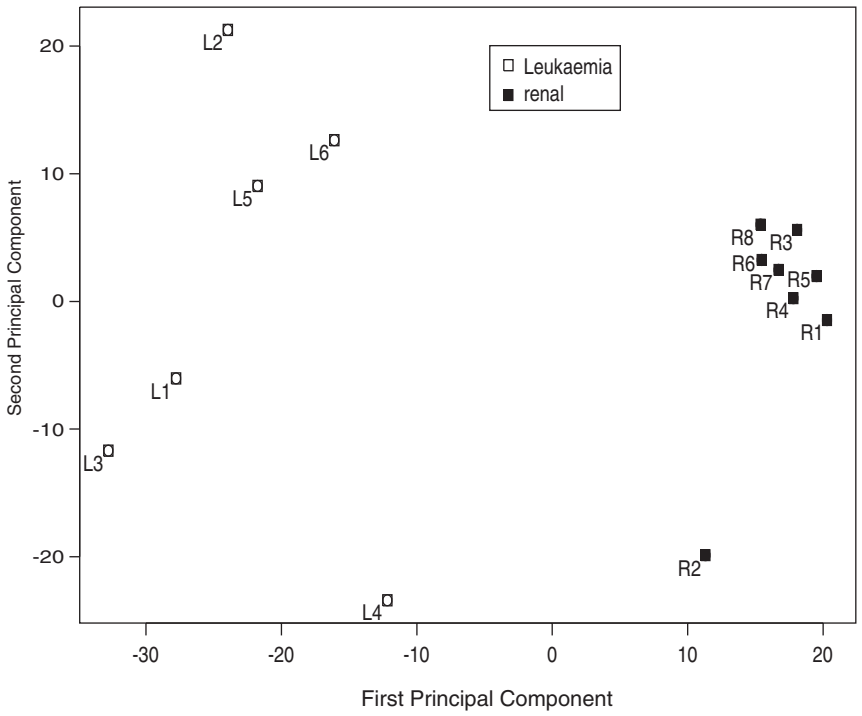
When q decreases, DIST_q^2 will increase and reach its maximum when $q=M$, hence R_q is a non-negative function of q. $R_q>1$ can be interpreted as: the reduction of distance between principal component scores for $q-1$ genes and q genes is larger than the average reduction of distance of including one more gene from the remaining $p-q$ gene list. We choose $q^\star$ as the largest q value satisfying $R_q>1.0$ and ignore the remaining genes when distance reduction as a result of including one more gene from this remaining gene set is deemed as being less important. This definition is similar to that of W in Reference [10], and we have also used 1.0 as a cut-off point.

5. AN EXAMPLE

The NCI60 database contains expression values of more than 9000 genes of 60 human cancer cell lines from nine types of cancer including leukaemia, renal, breast, etc. Gene expression levels are expressed as $-\log(\mathrm{ratio})$, where ratio = the red/green fluorescence ratio after computational balancing of two channels. Readers are referred to Scherf *et al.* [11] for more details. The data set has been posted online at http://discover.nci.nih.gov

One of the objectives is to explore the relationship between gene profiles and cancer phenotypes. In Reference [11], they have conducted a cluster analysis method to study the relationship. They finally retained 1375 genes and showed that most cell lines cluster together according to their phenotypes. Our approach will serve as a second step reduction method after Scherf's has been used.

For simplicity, we only study the cell lines from two types of cancer, leukaemia (six cell lines) and renal (eight cell lines); each cell line has microarray expression of the same 1375 genes. Using our proposed approach, we can select the most informative genes that

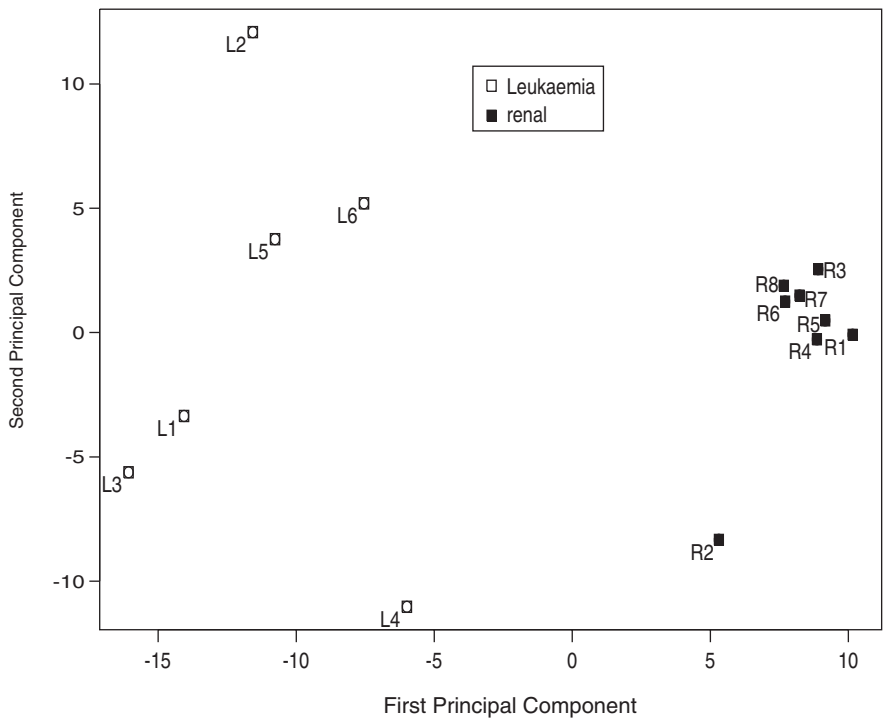


Plot 1. First two PCs of 1375 genes.

preserve the original data structure. Following the procedure we described in Section 2, we have found that the essential dimensionality of the original data set M is 8 (i.e. the first eight PCs account for over 80 per cent of the total variation of the original data set). Therefore, at least eight genes are needed in our selected gene subset to keep the original data structure.

Our gene selection procedure starts after we have determined the essential dimensionality M. Based on our gene selection procedure, the optimum number of genes (the number of genes which should be included in our gene subset) has been determined as $q^{\star} = 277$ for this data set. Once the top 277 genes have been selected using K's backward procedure, we can plot the first two principal components computed from all 277 genes and compare it with the corresponding plot for 1375 genes (see plot 1 and 2).

Interestingly, these two plots look quite similar. A referee has pointed out that the plots of first two PCs does not tell us about the first eight PCs. We present them here because the first two PCs account for the largest proportions of total variation, the closeness of the top eight PCs of the gene subset and the complete data set can at least be partially reflected by these plots. We can also measure the squared distance between the top eight principal components of our selected gene subset and the original data set by DIST^2. For this subset, DIST_K^2 (here the subscript K represents K's method) is equal to 4040.46. Also from plot 1 and 2, we find that the gene expression values of cell lines from different cancers are quite dissimilar. Six leukaemia cell lines (L1–L6) have different gene expression values from those

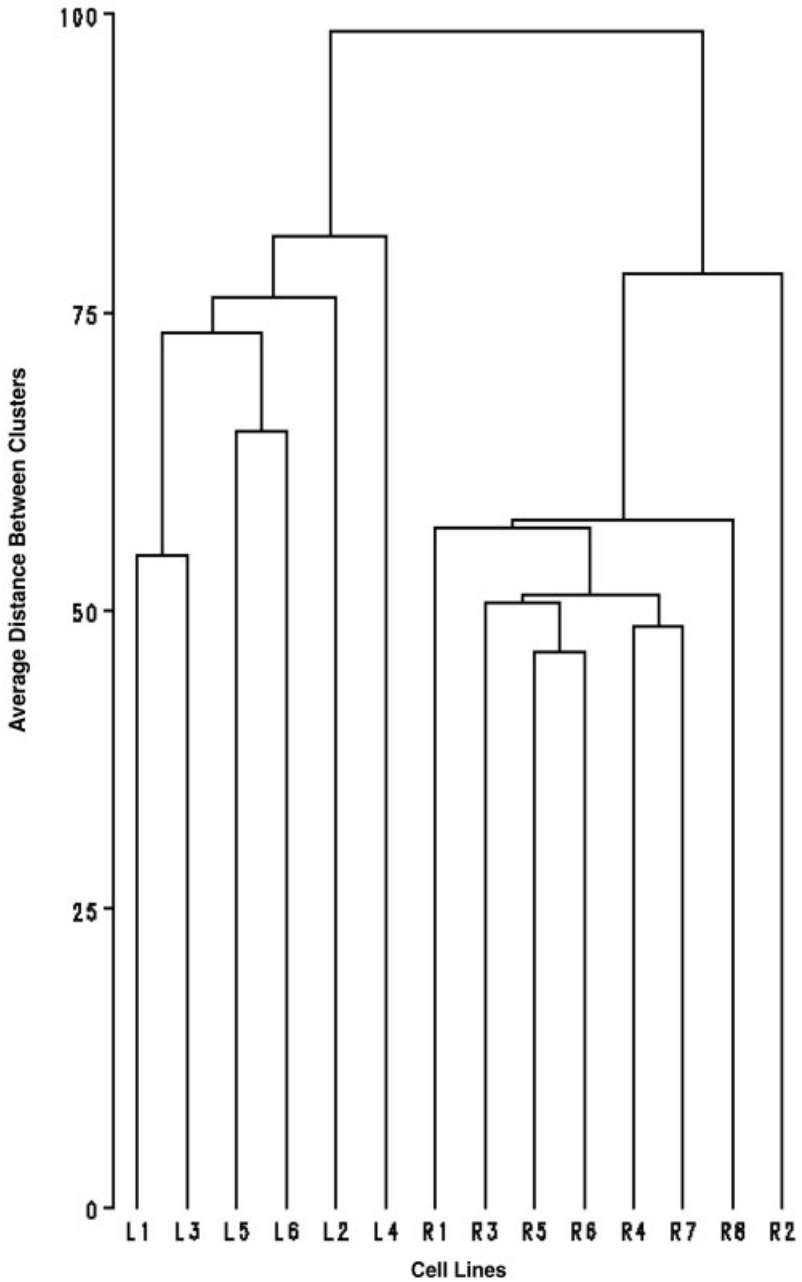


Plot 2. First two PCs of 277 genes selected by Krzanowski's method.

of the renal cell lines (R1 to R8). This fact can be further demonstrated by using 'Hierarchical Clustering'. Hierarchical clustering is a cluster analysis tool which results in groups of related samples that can be visualized with a dendrogram, a tree-based two-dimensional plot. Clusters are formed in an iterative manner using a bottom-up approach based on the distance measure of choice. Hierarchical clustering based on the average linkage approach is outlined in Reference [1] and can be applied to this data set using the SAS procedure CLUSTER. After performing the hierarchical clustering for both the original and the selected gene data sets, we have obtained two dendrograms: one based on all 1375 genes and the other one based on the selected 277 genes (dendrograms 1 and 2). As we have seen in our PC plots, two general clusters are emanating from these two dendrograms: one consisting of 6 leukaemia cell lines (L1–L6) and the other one consisting of 8 renal cell lines (R1 to R8). By comparing these two dendrograms, we find that these two dendrograms are quite similar.

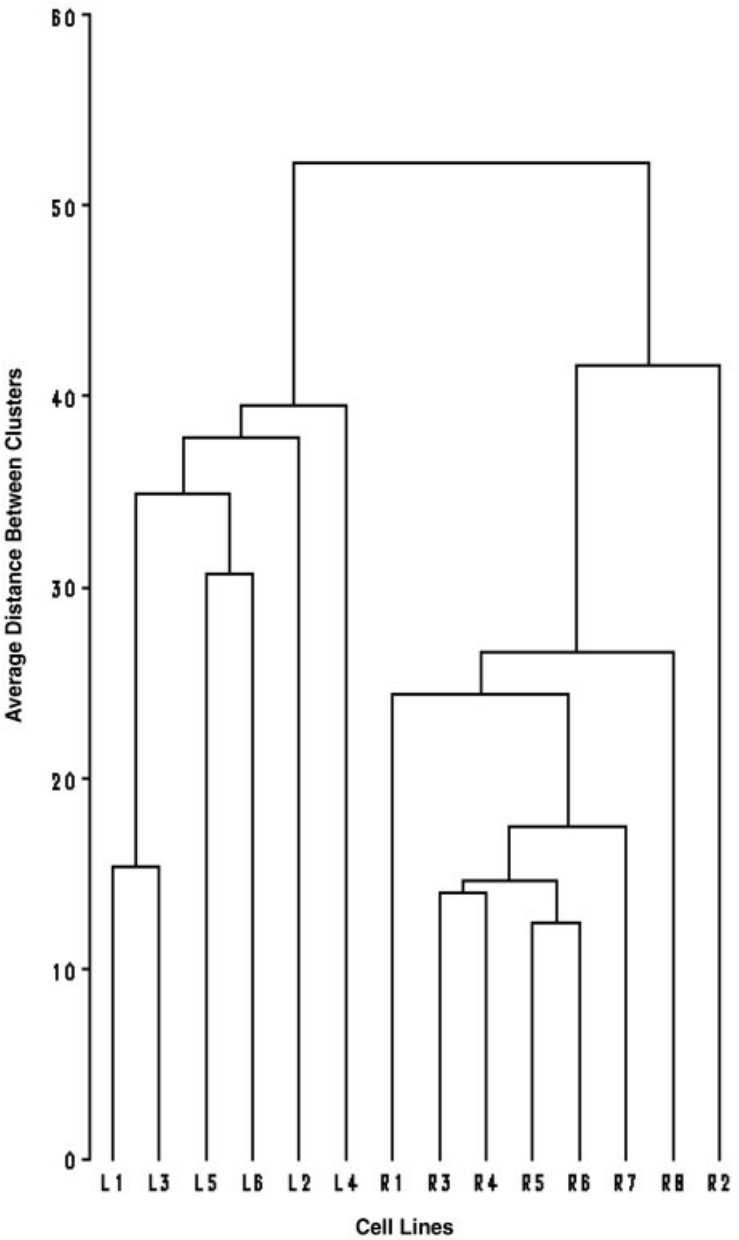
Based on previous comparisons, we conclude that our strategy works very well in selecting genes that can preserve the relationship between gene profiles and cancer phenotypes. The gene expression levels are quite different for these two types of cell lines.

If we apply Krzanowski's stopping rule [9] discussed in Section 4, we find that when $\mathrm{DIST}^2 = 76.35$ first exceeds 76.05: the corresponding 95 per cent critical value of the chi-squared stopping rule (this is the time to end the variable selection procedure according to K's stopping rule), our selected gene subset should contain 1130 genes. Based on 1130 genes



Dendrogram 1. Hierarchical clustering based on 1375 genes.

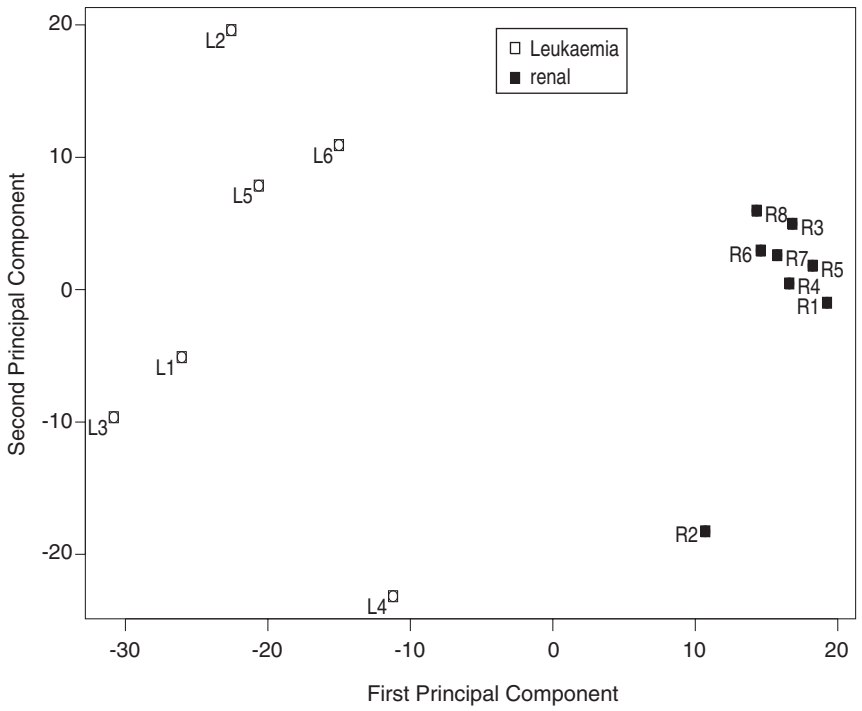
selected using K's method, we can obtain a plot of first two PCs (plot 3) and a dendrogram (dendrogram 3). By comparing the plot of PCs and the dendrogram with those based on 277 genes, we can see that they are very similar. Because 277 genes have kept the original data



Dendrogram 2. Hierarchical clustering based on 277 selected genes using Krzanowski's method.

structure almost as well as 1130 genes, we deduce that our stopping rule is quite effective in determining the right number of genes to preserve the original data structure.

To build effective classifiers based on those selected genes, we apply diagonal linear discriminant analysis. Diagonal linear discriminant analysis is a version of linear discriminant

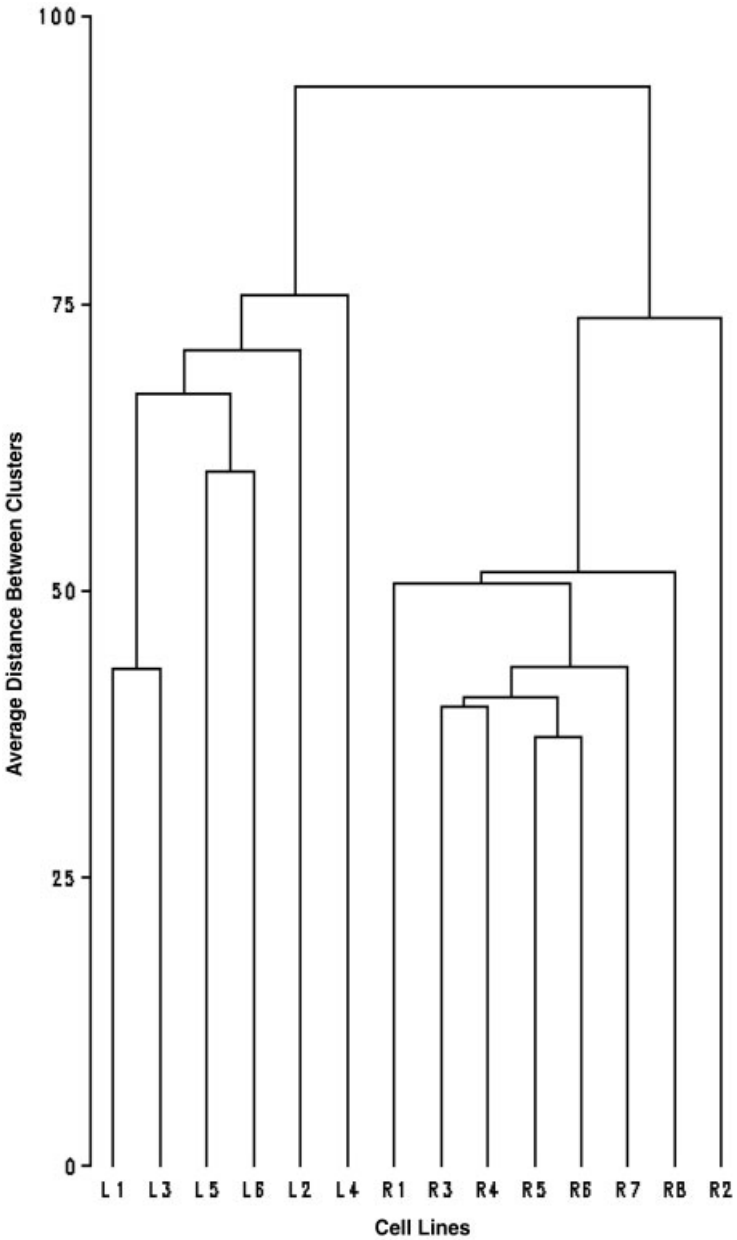


Plot 3. First two PCs of 1130 genes selected by Krzanowski's method.

analysis that ignores correlations among the genes in order to avoid over-fitting the data. The study by Dudoit *et al.* [12] has shown that diagonal linear discriminant analysis performs as well as much more complicated methods on varied microarray data sets. In order to evaluate the accuracy of this classifier based on these 277 genes, we use leave-one-out cross-validation approach. The cross-validation process omits one cell line at a time. For each cell line omitted, the entire analysis is repeated from scratch, including the determination of which genes are univariately significant on the reduced training sample. From that gene list, a multivariate predictor is constructed and applied to the sample that was omitted. These steps are repeated until we have finished omitting all of the samples one at a time. An overall cross-validated misclassification rate can be obtained after the cross-validation.

The software BRB-Array tools developed by Simon and Lam [13] offers both diagonal linear discriminant analysis and the cross-validation tools. We have used BRB-array tools to perform above analyses for our selected data and obtained the following results based on our selected genes: among these 277 genes, 69 are significant at $\alpha = 0.001$ level (based on two-sample t-test). A diagonal linear discriminant classifier built upon these 69 genes gives us a 0 per cent misclassification rate.

Similar analyses can also be performed for the original data set which contains 1375 genes, among all these genes, 233 have been identified significant at $\alpha = 0.001$ significant level. The same misclassification rate 0 per cent can be obtained if we use these 233 differentially expressed genes to build our classifier. Because we have only used 69 out of 277 significant



Dendrogram 3. Hierarchical clustering based on 1130 selected genes using Krzanowski's method.

genes to construct the classification rule which is as powerful as the classification rule constructed based upon 233 out of 1375 significant genes, we conclude that our gene subset keeps the original data structure quite well and the significant genes within this subset can be used to build an effective classifier to distinguish samples from different groups.

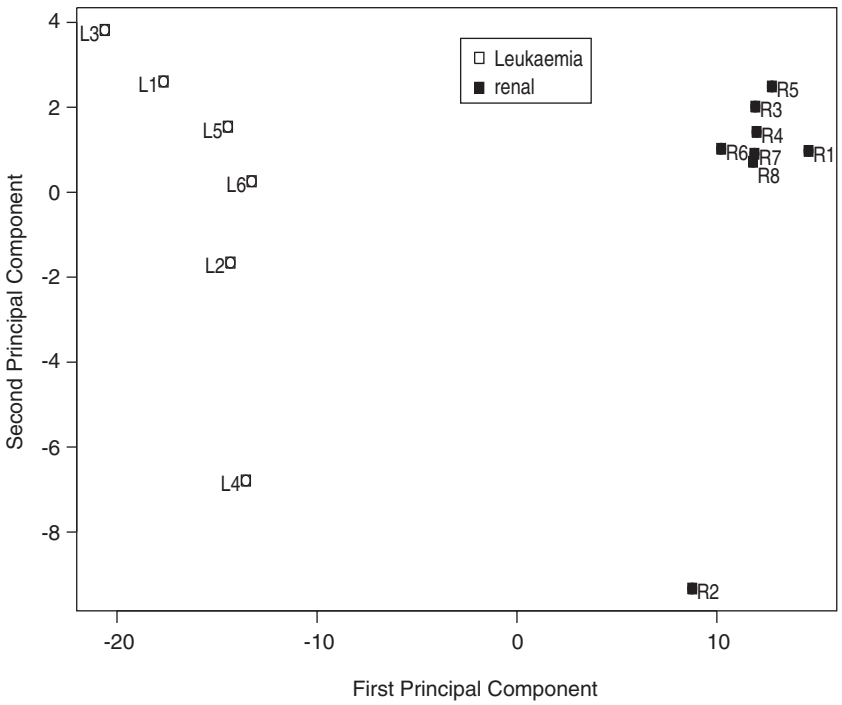
6. COMPARISONS WITH OTHER GENE SELECTION METHODS

As mentioned in Section 1, besides Krzanowski's procedure we just introduced, there are some other ways to perform the gene selection. Among them, the following two are the most popular and effective ones:

1. *t-test based method*: we first perform a two sample t-test for each gene (assuming there are two groups). The most important genes are those with smallest adjusted p-values. It has been noticed that the rankings of genes is invariant to p-value adjustment (see Reference [1]). Therefore, we can select the most important genes according to the rankings of corresponding unadjusted p-values.
2. *Jolliffe's method*: after determining the essential dimensionality M, we select genes (variables) with the largest absolute coefficients from each of the M leading principal components.

We have used the same NCI data set to perform the gene selection. To compare the performance of different gene selection procedures in a fair way, we retain the same number of genes (277) in our selected gene subset as we have done using K's method. If we use the t-test based method, after performing two-sample t-tests, we keep the top 277 genes according to the rankings of the raw p-values (Among them, the top 99 genes are significant in terms of adjusted p-values. We expect that the subset including 277 genes can recover the original data structure better than the one including only 99 genes). If we use Jolliffe's method, because the essential dimensionality is 8, we can also choose the top 277 genes from the eight leading principal components. As a result, approximately 35 different genes with largest coefficients in each of the eight principal components have been selected.

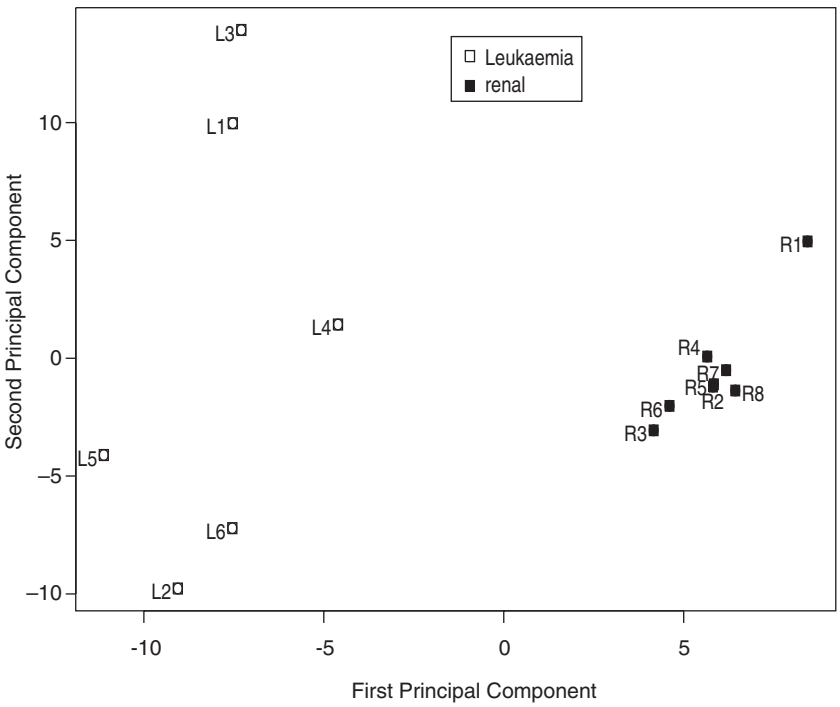
Based on selected gene subsets using these two alternative methods, we obtain the plots of first two PCs (plots 4 and 5) and the corresponding dendrograms (dendrograms 4 and 5). By comparing plot 4 to plot 1 and plot 2, we find that a better separation can be achieved between the leukaemia cell lines and renal cell lines if we apply the t-test based approach. This is not surprising because as we mentioned before, the t-test based approach tends to overstate the dissimilarity between gene expression values. We have also found that the original data structure has not been kept very well using the t-test based method (especially for the leukaemia cell lines): the relative locations of different cell lines have changed much more dramatically than those in plot 2 (all compared to the relative locations in plot 1). In fact, for this gene subset we have $\mathrm{DIST}^2 = 5597.97$ which is much larger than the squared distance between the PCs of the original and the selected gene subset if we use K's method ($\mathrm{DIST}_K^2 = 4040.46$). If we compare dendrogram 4 to dendrograms 1 and 2, we can see that in all these dendrograms, there are two general clusters: one consisting of the leukaemia cell lines (L1–L6), and the second consisting of renal cell lines (R1–R8). In dendrograms 1 and 2, within the leukaemia cluster, L4 is separated from other cell lines, L2 and the branch formed by L1, L3, L5 and L6 are also separated; within the renal cluster, R2 is separated from other cell lines, R8 is separated from the branch formed by R1, R3, R4, R5, R6 and R7. In dendrogram 4, these relationships have not been preserved very well, for example, within the leukaemia cluster, L4 has not been separated from other cell lines; within the renal cluster, although R2 is separated from other cell lines, the difference between R8 and the branch formed by R1, R3, R4, R5, R6 and R7 has not been identified either.



Plot 4. First two PCs of 277 genes selected by *t*-test based method.

Similarly, if we compare plot 5 to plots 1 and 2, we can also find that the relative locations of different cell lines in plot 5 have changed more dramatically than those in plot 2 (compared to plot 1). $\mathrm{DIST}^2 = 4688.75 > \mathrm{DIST}_K^2 = 4040.46$. If we explore the corresponding dendrograms (dendrogram 5 to dendrograms 1 and 2), we also find that dendrogram 5 has not recovered relationships between cell lines as well as dendrogram 2. The obvious difference between dendrogram 5 and dendrogram 1 or 2 is that in dendrogram 5, the leukaemia cell lines and the renal cell lines have not been separated in different clusters. This result contradicts the fact that has been shown in both dendrograms 1 and 2.

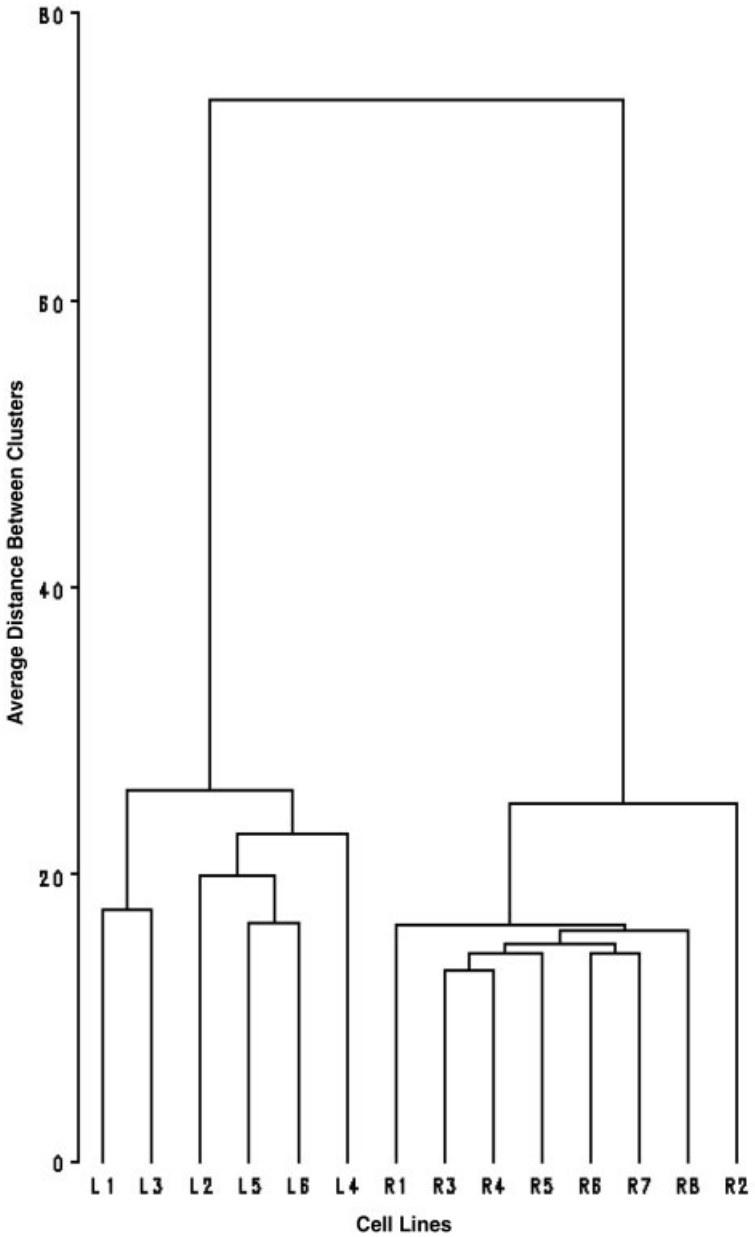
Surprisingly, if we compare plot 5 to dendrogram 5, we have found some discrepancies between these two figures: for example, from plot 5, it seems that L1, L3 should be separated from renal cell lines; while in dendrogram 5, they have been clustered together. This 'strange' phenomenon happens because these two types of figures are not equivalent: the dendrogram is based on the distance between pairs of samples (the distance measure is calculated using the original gene expression vectors of two different samples, see Reference [1]) while the distance in plot 5 is calculated based on the first two PCs which are the linear combination of the original genes (variables). Furthermore, as pointed out by the referee, the plot of first two PCs does not provide us with all the information about the top 8 PCs. To show this, we have produced plot 6. From the plot, we can see that L1, L3 are mixed together with the renal cluster, R2 seems to be quite different from other renal cell lines and R8 is separated from the renal cluster. These facts can be used to explain the results



Plot 5. First two PCs of 277 genes selected by Jolliffe's method.

in dendrogram 5. We can therefore reach a conclusion that although the plots of first 2 or 3 PCs can help us interpret the relationships between cell lines intuitively sometimes, we should not just rely on them to explain the clustering results, especially when M is much larger than 2 or 3. Exactly the same arguments can be used to explain other discrepancies between dendrograms and plots of first two PCs: for example, if we compare the plots 1–3 to corresponding dendrograms, we find that in all these plots, R8 is not separated from the branch formed by R1, R3, R4, R5, R6 and R7; while in dendrograms 1–3, they are separated. We have also produced plot 7 to show the relative locations of R8 and the branch formed by R1, R3, R4, R5, R6 and R7 in terms of the third and fourth PCs of the original data set. The difference between them can help us explain the results in dendrogram 1. If we plot the 3rd and 4th PCs for the selected 277 or 1130 genes, similar difference can be detected.

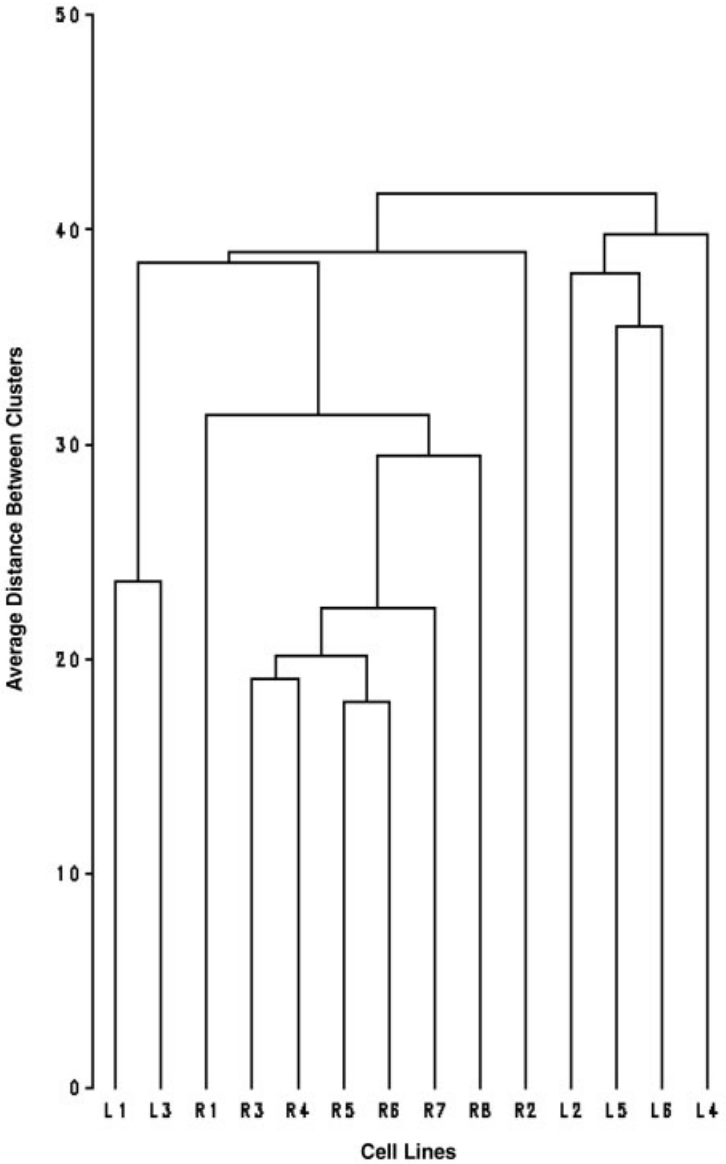
Overall, based on previous comparisons with respect to plots of PCs, the squared distance between corresponding principal components and dendrograms, we can conclude that the gene subset chosen by either the t-test based method or Jolliffe's method has not kept the original data structure as well as the subset selected by K's approach. Although the t-test based approach is able to identify genes that can distinguish leukaemia and renal cell lines, these genes do not preserve the original relationships between cell lines very well. The subset determined by Jolliffe's method does not recover these relationships very well either.



Dendrogram 4. Hierarchical clustering based on top 277 genes using two sample *t*-test.

7. DISCUSSION

Our proposed gene selection approach is built upon Krzanowski's variable selection strategy [5]. Basically, we have extended his idea of variable selection to microarray data analysis.

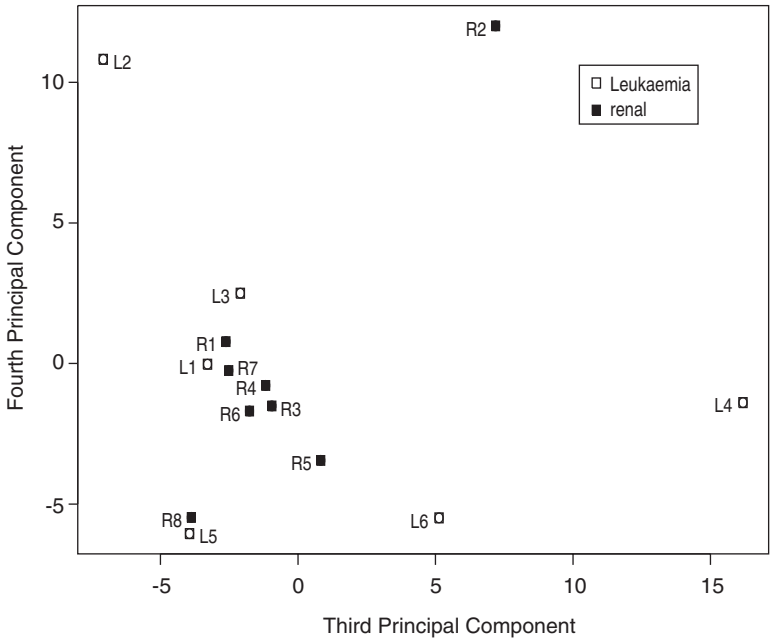


Dendogram 5. Hierarchical clustering based on 277 selected genes using Jolliffe's method.

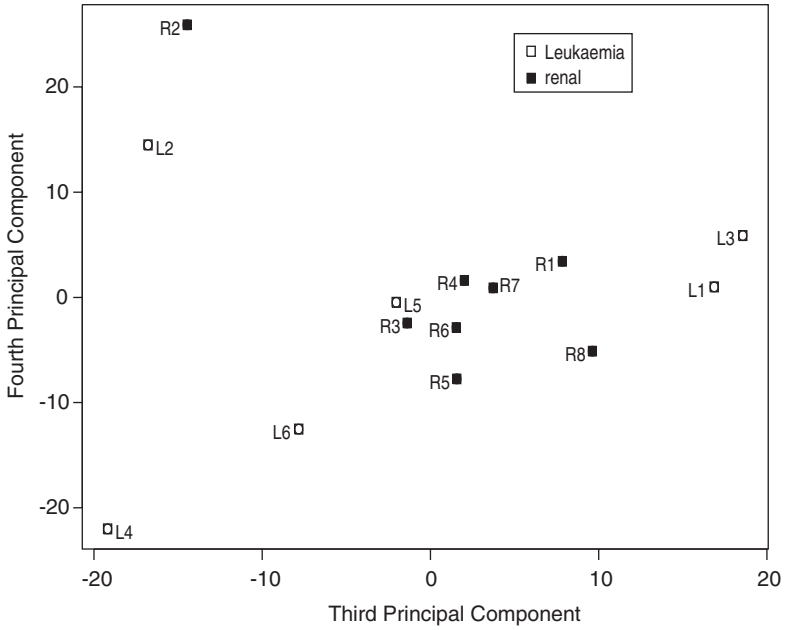
Because we have not imposed any restrictions on group labels when we perform the gene selection, this strategy can be applied for both the supervised and the unsupervised study.

As demonstrated in Section 5, our gene selection approach is powerful in recovering the original data structure, comparing gene expression values of samples from different groups and building effective classifiers.

By comparing the proposed method to *t*-test based method and Jolliffe's method, we find that Krzanowski's method selects the gene subset that can preserve the original data structure



Plot 6. The 3rd and 4th PCs of 277 genes selected by Jolliffe's method.



Plot 7. The 3rd and 4th PCs of 1375 genes.

more effectively. Hence the proposed method can be applied to select the most informative genes that can explain clinically related differences between individual subjects more easily and clearly. Extension of this strategy to gene expression comparison of samples from more than two groups is also straightforward.

ACKNOWLEDGEMENTS

The authors thank two referees for a thorough review and constructive comments that greatly improved the paper. The authors especially thank Dr Aiyi Liu for very valuable discussions.

REFERENCES

1. Satagopan JM, Panageas KS. Tutorial in biostatistics: a statistical perspective on gene expression data analysis. *Statistics in Medicine* 2003; **22**:481–499.
2. Jolliffe IT. Discarding variables in a principal component analysis I: artificial data. *Applied Statistics* 1972; **21**:373–374.
3. Jolliffe IT. Discarding variables in a principal component analysis II: real data. *Applied Statistics* 1973; **22**: 21–31.
4. Liu A, Zhang Y, Gehan E, Clark R. Block principal component analysis to gene microarray data classification. *Statistics in Medicine* 2002; **21**:3465–3474.
5. Krzanowski WJ. Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics* 1987; **36**:22–33.
6. Jolliffe IT. *Principal Component Analysis* (2nd edn). Springer: New York, 2002.
7. Sibson R. Studies in the robustness of multidimensional scaling: procrustes statistics. *Journal of Royal Statistical Society, Series B* 1978; **40**(2):234–238.
8. Gower JC. Statistical methods of comparing different multivariate analyses of the same data. In *Mathematics in the Archaeological and Historical Sciences*, Hodson FR, Kendall DG, Tauta P (eds). University Press: Edinburgh, 1971; 138–149.
9. Krzanowski WJ. A stopping rule for structure-preserving variable selection. *Statistics and Computing* 1996; **6**:51–56.
10. Krzanowski WJ. Cross-validation in principal component analysis. *Biometrics* 1987; **43**:575–584.
11. Scherf W, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Bostein D, Brown PO, Weinstein JN. A gene expression data base for the molecular pharmacology of cancer. *Nature Genetics* 2000; **24**:236–244.
12. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association* 2002; **97**:77–87.
13. Simon RM, Lam A. *BRB - Array Tools User's Manual (Version 3.2)*. Technical Report 007, Biometric Research Branch, National Cancer Institute 2003.
14. Cadima JFCL, Jolliffe IT. Variable selection and the interpretation of principal subspaces. *Journal of Agricultural, Biological and Environmental Statistics* 2001; **6**:62–79.
15. Eastment HT, Krzanowski WJ. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* 1982; **24**:73–77.
16. Khattree R, Dayanand Naik DN. *Multivariate Data Reduction and Discrimination with SAS Software*. SAS Institute Inc.: Cary, NC, 2000.
17. Wright GW, Simon RM. The randomized variance model for finding differentially expressed genes. *Technical Report* xx, Biometric Research Branch, National Cancer Institute, 2003.

