SamCluster: an integrated scheme for automatic discovery of sample classes using gene expression profile

Wuju Li1,*, Ming Fan1 and Momiao Xiong2

1Beijing Institute of Basic Medical Sciences, PO Box 130(3), Beijing 100850, People's Republic of China and 2Human Genetics Center, School of Public Health, University of Texas-Houston Health Science Center, TX 77225, USA

Received on April 17, 2002; revised on October 27, 2002; November 11, 2002; accepted on November 13, 2002

ABSTRACT
Motivation: Feature (gene) selection can dramatically improve the accuracy of gene expression profile based sample class prediction. Many statistical methods for feature (gene) selection such as stepwise optimization and Monte Carlo simulation have been developed for tissue sample classification. In contrast to class prediction, few statistical and computational methods for feature selection have been applied to clustering algorithms for pattern discovery.

Results: An integrated scheme and corresponding program SamCluster for automatic discovery of sample classes based on gene expression profile is presented in this report. The scheme incorporates the feature selection algorithms based on the calculation of $CV$ (coefficient of variation) and $t$-test into hierarchical clustering and proceeds as follows. At first, the genes with their $CV$ greater than the pre-specified threshold are selected for cluster analysis, which results in two putative sample classes. Then, significantly differentially expressed genes in the two putative sample classes with $p$-values $\leq 0.01$, $0.05$, or $0.1$ from $t$-test are selected for further cluster analysis. The above processes were iterated until the two stable sample classes were found. Finally, the consensus sample classes are constructed from the putative classes that are derived from the different $CV$ thresholds, and the best putative sample classes that have the minimum distance between the consensus classes and the putative classes are identified. To evaluate the performance of the feature selection for cluster analysis, the proposed scheme was applied to four expression datasets COLON, LEUKEMIA72, LEUKEMIA38, and OVARIAN. The results show that there are only 5, 1, 0, and 0 samples that have been misclassified, respectively. We conclude that the proposed scheme, SamCluster, is an efficient method for discovery of sample classes using gene expression profile.

Availability: The related program SamCluster is available upon request or from the web page http://www.sph.uth.tmc.edu:8052/hgc/Downloads.asp.

Contact: wujuli@yahoo.com

INTRODUCTION
Microarray technology provides a powerful tool for simultaneous analysis of expression of thousands of genes (Alizadeh et al., 2000). A popular application of microarray is to cluster tissue sample or group genes, which can be used to detect subtypes of diseases, such as tumor subtype, to investigate function of genes, and to produce therapies. Many clustering methods such as hierarchical clustering (Eisen et al., 1998), self-organizing map (SOM) (Tamayo et al., 1999), and $k$-means (Tavazoie et al., 1999) have been applied to discovering the sample class pattern hidden in the expression datasets. Cluster analysis shows high power of expression profiles for discriminating the types of samples. However, unlike sample class prediction, which uses feature selection as a tool to select a small number of genes for classifying types of samples (Li and Xiong, 2002), cluster analysis uses a large number of genes for identifying sample classes and has not explored gene selection for data reduction. It is noted that even if using the same clustering algorithm, we may get different sample dendrograms for different gene subsets. Therefore, gene selection plays an important role in sample class discovery too.

Recently, some methods for gene selection in the cluster analysis have been proposed. For example, Lukashin and Fuchs (2001) used the following criteria to select genes for clustering yeast cell cycle data if their absolute values of expression at all 17 time points are equal to or greater than 100 or they have at least a 2.5-fold change.

*To whom correspondence should be addressed.
in expression level during the time-course. In addition, Welsh et al. (2001) used SD ≥ 250, an option provided by the program CLUSTER (Eisen et al., 1998), as the criterion to select 1243 genes for clustering the epithelial ovarian cancer dataset. Perou et al. (1999) used R/G ratios that varied 3-fold or more in three or more of the 26 samples/experiments as the criterion to select 1247 genes for clustering 26 samples.

However, these criteria for gene selection are problem-dependent. They may only be effective for some particular expression dataset, but may not be so effective for other expression datasets. Recently, Xing and Karp (2001) presented a unified scheme for sample class discovery based on normalized cuts. They further took the LEUKEMIA72 as an example to test the performance of the scheme. But the number of features for reference partitions was chosen empirically and the clustering result is sensitive to the number of selected features. In order to solve above problems, we presented an integrated scheme for automatic discovery of sample classes, which incorporates feature selection based on the calculation of CV, t-test into hierarchical clustering and systematically determine the optimal number of features for sample class discovery. Furthermore, by introducing the concept of consensus sample classes, we improved the robustness of the proposed scheme for sample class identification. The scheme was evaluated using four gene expression datasets, namely, COLON (Alon et al., 1999), LEUKEMIA72 and LEUKEMIA38 (Golub et al., 1999), and OVARIAN ((Welsh et al., 2001).

**MATERIALS AND METHODS**

**Gene expression datasets**

Four gene expression datasets from three expression profiles were used. The first dataset (COLON) consists of expression profiles of 2000 genes from 22 normal tissues and 40 tumors and was retrieved from http://www.molbio.princeton.edu/colondata (Alon et al., 1999). The second dataset (LEUKEMIA72) consists of expression profiles of 6817 genes from 47 type I leukemias (called ALL) and 25 type II leukemias (called AML) and was retrieved from http://www.wi.mit.edu/MPR/data_set_ALL_AML.html (Golub et al., 1999). In addition, a sub-dataset (LEUKEMIA38: 27 ALL and 11 AML) was extracted from LEUKEMIA72, which was used to compare the integrated scheme with the Golub’s SOM method. The fourth dataset OVARIAN consists of expression profiles of 7129 genes from 27 epithelial ovarian cancer, five normal tissues, four malignant epithelial ovarian cell lines which was retrieved from http://www.gnf.org/cancer/ovary (Welsh et al., 2001).

**Calculation of coefficient of variation (CV)**

Let \( G = [g_{ij}]_{m \times n} \) stands for gene expression matrix where each row represents the expression level of a gene in the samples and each column represents the observed expression profiles of all genes in a sample. \( m \) and \( n \) represent the total number of genes and samples respectively. For the \( i \)th gene expression vector \( g_i = [g_{i1}, g_{i2}, \ldots, g_{in}] \), the coefficient of variation (CV) was defined as

\[
CV_i = \frac{S_i}{\bar{g}_i} \quad \text{where} \quad \bar{g}_i = \frac{1}{n} \sum_{j=1}^{n} g_{ij} \quad \text{and} \quad S_i = \frac{1}{n-1} \sum_{j=1}^{n} (g_{ij} - \bar{g}_i)^2.
\]

Clearly, \( CV_i \) is a quantity to measure the variability of expression level for each gene across all samples. A high \( CV_i \) indicates relative high variation of expression levels of the gene to the mean. Therefore, by intuition, only those genes with high \( CV_i \) values are of great potential to class discovery. During calculation of \( CV_i \)s, two special considerations were given. First, if the minimum value \( (g_{i\min}) \) of the particular gene expression vector is negative, the absolute value of \( g_{i\min} \) will be added into all the elements in this gene expression vector so that every element in gene vector is greater than or equal to zero. Second, in order to avoid the bigger effect of the maximum and minimum values of the gene expression vector on the CV value, all the maximum and minimum values of the gene expression vector have been discarded before the calculation of \( CV_i \), but they are kept for the hierarchical cluster analysis. Finally, Let \( CV = [cv_1, cv_2, \ldots, cv_m]' \) stand for the vector of coefficient of variation for all genes, and \( MCV \) and \( SCV \) stand for the average value and standard deviation of the \( CV_i \)s.

**T-test**

Standard t-test statistic was used to test for the differential expressions of the gene in two identified putative sample classes. In this stage, the original gene expression data were used.

**Short description of hierarchical clustering analysis**

Hierarchical clustering analysis was carried out by average linkage method. We assume that the total number of samples is \( n \). We begin with calculation of the standard correlation coefficient matrix of the selected set of genes that will be used for measuring the similarity between pairs of samples. Then, we find the highest correlation between two samples (the most similar pair of samples) by searching the correlation coefficient matrix and merge the corresponding samples, to get a cluster denoted by a node, says \( S_A \), in a dendrogram. The correlation coefficient between the new formed cluster \( (S_A) \) and other cluster (for example, \( S_B \)) is calculated by taking the average of the correlation coefficients between members in the cluster \( S_A \) and members in the cluster \( S_B \). The correlation coefficient matrix is updated by (1) deleting the rows and columns...
corresponding to cluster $S_A$ and (2) adding a row and column representing the correlation coefficient between $S_A$ and the remaining clusters. The above processes (1) and (2) are repeated for $n-1$ times until all $n$ samples are clustered in the single dendrogram (Eisen et al., 1998).

**Integrated scheme for automatic discovery of sample classes and program design**

Integrated scheme for sample class discovery consists of several steps. First, the genes with their coefficient of variation greater than or equal to the given threshold $CV_{ih}$, which is defined as $CV_{ih} = M_{CV} + Ci \times S_{CV}$, where $Ci = 0.0, 0.1, \ldots, 2.0$ with 0.1 as the increment, are selected for clustering samples. Second, the cluster analysis is performed, which results in two putative sample clusters. But not all selected genes are differentially expressed in the two putative sample clusters. The genes showing no differential expressions will become sources of noise for the sample class discovery. To eliminate the genes with no differential expression, the $t$-test is then used to test for differential expression of the gene. The differentially expressed genes with low $P$ values (for example, $P = 0.01, 0.5,$ or $0.1$) are selected for the further cluster analysis. The above processes were iterated until two stable putative sample clusters were found. Obviously, different combination of threshold $CV_{ih}$ and $P$ values may give different final sample classes, which form a set of putative sample classes.

If the types of tissue sample are unknown, it is very difficult to assess which sample class pattern in the set of putative sample classes is the best. To solve this problem, a relationship matrix $S_{n \times n}$ for $n$ tissue samples was introduced. (We only consider two groups.) Initially, all elements in $S$ are assigned zero. For all combination of threshold $CV_{ih}(= 0.0, 0.1, \ldots, 2.0)$ and particular $P$ value ($= 0.01, 0.05,$ or $0.1$), $S(i, j) = S(i, j) + \delta_{ij}$ where $\delta_{ij} = 0$ if tissue $i$ and tissue $j$ are in the same putative sample class; $\delta_{ij} = 1$, otherwise. For any particular $P$ value ($= 0.01, 0.05,$ or $0.1$), we will obtain a relation matrix $S_{n \times n}$, which can be used to cluster tissue samples again. We call the putative sample clusters found from $S_{n \times n}$ as the consensus sample clusters. Finally, the distances between the consensus sample clusters and each putative sample clusters from every combination of threshold $CV_{ih}$ and $P$ value were calculated and those putative sample clusters with the minimum distances were selected as the best putative sample classes.

Based on the integrated scheme proposed here, we have developed a program SamCluster in GUI mode in Matlab language. The gene expression matrix, which was stored in a text file, was input into the program through the standard file dialog box. The first row and column of the file represent the sample names and gene names respectively. The other elements stand for the gene expression level. Subsequently, the best putative sample classes were found automatically. The detailed information about the usage of the program is stored in a help file. Figure 1 is the demonstration of the interface.

**RESULTS**

**Sample class discovery without gene selection and iteration**

To explain the importance of gene selection, three datasets COLON, LEUKEMIA72, and LEUKEMIA38 were clustered directly without gene selection. Two putative sample classes were extracted. The detailed class information is given in Table 1.

From Table 1, it can be seen that the true sample class pattern cannot be found without gene selection.

**Sample class discovery using the integrated scheme**

To explore the effects of different combination of $CV$ thresholds and $P$ values on the final sample class pattern, different $CV$ thresholds and $P$ values were utilized. For $CV$ threshold, it is reasonable to assume that only those genes with $CV_{ih} \geq M_{CV}$ are informative to the sample class discovery. Therefore, the $CV_{ih}$ takes the following form: $CV_{ih} = M_{CV} + CVi \times S_{CV}$, where $CVi = 0.0, 0.1, \ldots, 2.0$ with the increment by 0.1. Meanwhile, the following $P$ values 0.01, 0.05, and 0.1

![Fig. 1. The main interface of the program SamCluster.](Image)

<table>
<thead>
<tr>
<th>Class</th>
<th>COLON</th>
<th>LEUKEMIA72</th>
<th>LEUKEMIA38</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>14 Normal + 28 Tumor</td>
<td>18 ALL + 18 AML</td>
<td>8 ALL + 10AML</td>
</tr>
<tr>
<td>Class II</td>
<td>8 Normal + 12 Tumor</td>
<td>29 ALL + 7 AML</td>
<td>19 ALL + 1 AML</td>
</tr>
</tbody>
</table>

Table 1. Detailed class information for datasets: COLON, LEUKEMIA72, and LEUKEMIA38
were used respectively in \( t \)-test for gene selection. The relationship between the number of misclassified samples and each of different combinations of \( CV \) thresholds and \( P \) values was given in Table 2. From Table 2, it can be seen clearly that the minimal number of misclassified samples for COLON, LEUKEMIA72 and LEUKEMIA38 is 5, 1 and 0 respectively according to the comparison of the putative sample classes to the sample pathological phenotype. Different combination of \( CV \) thresholds and \( P \) values generated different putative sample classes. When the information about sample pathological phenotype is not available, it is very difficult to determine which combination is the best. To overcome the problem, following two steps were taken. The first step is to construct the consensus sample classes from the inferred putative sample classes for different \( CV \) thresholds and \( P \) values. The second step is to calculate the distance between the consensus sample classes and each putative sample classes identified by each combination of parameters. Finally, the best putative sample classes with the minimal distance were found. All distance information was also provided in D columns in Table 2.

From the D columns in Table 2, it can be seen that the best putative sample classes for COLON with the correct rate 88.7% (55/62), 90.3% (56/62), and 90.3% (56/62) respectively were derived from \( CV = 5, 0.1, 1 \). Up to now, many cluster methods have been used to cluster the dataset COLON. For example, Alon et al. (1999) used the two-way clustering method to search the class pattern based on the 20 genes with the most significantly differential expressions between tumors and normal tissues. Two distinct clusters were found with five tumors clustered with normal tissues and three normal tissues clustered with tumors. The accurate rate is 87.1% (54/62). The classification accuracy rate based on the Fisher’s linear discriminate method and two selected two genes can reach 92.0% (57/62) (Xiong et al., 2001). However, the Fisher’s linear discriminate method assumed that there exist two classes, i.e. tumors and normal tissues, even before any cluster analysis. Our

was given in Table 2. From Table 2, it can be seen clearly that the minimal number of misclassified samples for COLON, LEUKEMIA72 and LEUKEMIA38 is 5, 1 and 0 respectively according to the comparison of the putative sample classes to the sample pathological phenotype. Different combination of \( CV \) thresholds and \( P \) values generated different putative sample classes. When the information about sample pathological phenotype is not available, it is very difficult to determine which combination is the best. To overcome the problem, following two steps were taken. The first step is to construct the consensus sample classes from the inferred putative sample classes for different \( CV \) thresholds and \( P \) values. The second step is to calculate the distance between the consensus sample classes and each putative sample classes identified by each combination of parameters. Finally, the best putative sample classes with the minimal distance were found. All distance information was also provided in D columns in Table 2.

From the D columns in Table 2, it can be seen that the best putative sample classes for COLON with the correct rate 88.7% (55/62), 90.3% (56/62), and 90.3% (56/62) respectively were derived from \( CV = 5, 0.1, 1 \). Up to now, many cluster methods have been used to cluster the dataset COLON. For example, Alon et al. (1999) used the two-way clustering method to search the class pattern based on the 20 genes with the most significantly differential expressions between tumors and normal tissues. Two distinct clusters were found with five tumors clustered with normal tissues and three normal tissues clustered with tumors. The accurate rate is 87.1% (54/62). The classification accuracy rate based on the Fisher’s linear discriminate method and two selected two genes can reach 92.0% (57/62) (Xiong et al., 2001). However, the Fisher’s linear discriminate method assumed that there exist two classes, i.e. tumors and normal tissues, even before any cluster analysis. Our
scheme does not need this assumption. Based on gene expression matrix only, we can automatically find the putative sample classes. The correct rate reaches 90.3% (56/62), which is near to the Xiong’s results. In case that the sample phenotype is available, we can also get the same correct rate as Xiong’s results for the $CV_i = 0.5$ ($P = 0.01$), $CV_i = 1.0, 1.1, 1.2$, and $1.3$ ($P = 0.05$), and $CV_i = 0.5, 0.7, 0.8, 1.1, 1.2,$ and $1.3$ ($P = 0.1$) respectively.

For LEUKEMIA72, the best putative sample classes are derived from the $CV_i = 0.3, 0.4, 0.5, 0.6$ ($P = 0.01$), $CV_i = 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1.1$ ($P = 0.05$), and $CV_i = 0.2, 0.3, 0.5, 0.6, 0.7$ ($P = 0.1$) respectively. These combinations provide the same correct rate 97.2% (70/72). There are only two ALLs misclassified in the best putative sample classes. If we combine the cluster results and sample pathological phenotype, there is only one AML misclassified ($CV_i = 0.1, 0.2, 0.7, 0.8, 1.0$ for $P = 0.01$, $CV_i = 0.1$ for $P = 0.05$, and $CV_i = 0.0$ and 0.1 for $P = 0.1$). But the number of misclassified samples in the best putative sample classes identified by CLIFF is three (three ALLs are misclassified into the AML class, Xing and Karp, 2001). Based on the result presented here, our scheme clearly outperforms the Xing’s method.

For LEUKEMIA38, the best putative sample classes are derived from the $CV_i = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$ ($P = 0.01$), $CV_i = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3$ ($P = 0.05$), and $CV_i = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3$ ($P = 0.1$) respectively. These combinations also provide the same correct rate 94.7% (36/38). Compared to the Golub’s results where two classes were found using SOM method and four samples (one AML, three ALLs) were misclassified.
With respect to distance $D$ distribution, the best or near best putative sample classes can be found when $CV_i$ is in the distance $[0.5, 1.0]$. Relative to the sample class discovery, if the $CV_i$ is too low, some noise genes will be included; if the $CV_i$ is too high, a lot of signal genes will be discarded. In addition, the consensus sample classes from different $P$ values (0.01, 0.05 or 0.1) often give consistent results, and the correct rate is higher for $P = 0.05$ and $P = 0.1$ than for $P = 0.01$. Thus, for a particular gene expression profile, the above three $P$ (especially $P = 0.05$ or 0.1) values can be used to identify the best putative sample classes.

**Sample class discovery for OVARIAN dataset**

To further demonstrate the performance of the above scheme, the dataset OVARIAN was used. By inputting the gene expression matrix into the program SamCluster, the consensus sample classes were found automatically. The results indicated that there are three samples misclassified for $P = 0.01$. But for $P = 0.05$ and $P = 0.1$, the 5 normal tissues and 31 tumors (including four malignant epithelial ovarian cell lines) are separated perfectly. The dendrogram was displayed in Figure 3, which shows that our result performs better than the Welch’s results, where five normal tissues, four malignant epithelial ovarian cell lines, and 27 tumors are mixed together.

**DISCUSSIONS**

In this report, we present an integrated scheme combining $CV$ calculation, $t$-test and hierarchical clustering for automatic discovery of sample classes. Four gene expression datasets were used to evaluate the proposed scheme. The results indicated that the automatically identified best putative classes were basically consistent with the sample phenotypes using the same scheme. Compared to the CLIFF program (Xing and Karp, 2001), both methods share some common features. First, the genes were selected in unsupervised way to obtain the initial reference partition of samples. Then, the genes were selected in supervised mode based on the reference partition. Finally, iterative processes were used to generate the final stable putative sample classes. The main differences between SamCluster and CLIFF are as follows. First, SamCluster used the hierarchical clustering analysis, but CLIFF used the normalized cut clustering method. Meanwhile, we also explored the effects of the $K$-means and SOM methods by replacing the hierarchical clustering method with them in the integrated scheme. The results were not comparable to those from the hierarchical clustering (data not shown). Second, in unsupervised stage, SamCluster used $CV$ to select genes for the initial reference partition of samples. There is no necessary in our scheme to assume that the gene expression profile abide by the Gaussian distribution. The initial reference partition was obtained using those genes with their $CV_i$ greater than or equal to the given threshold. CLIFF assumes that each feature comes from a Gaussian distribution. Then, CLIFF models the likelihood of each feature as a univariate mixture of two Gaussians and ranks all features (genes) by Bayes errors. Finally, the initial reference partition was obtained using the certain number of genes with the smallest Bayes errors. Both methods have some difficulty in determining the optimal number of genes for the initial reference partition. Third, in supervised stage, SamCluster takes $t$-test to select those genes with significantly differential expressions for iterations. CLIFF takes both ‘information gain ranking’ and ‘Markov blanket filtering’ methods to select the certain number of genes with the smallest $I_{gain}$ values or $\Delta (F_i | M)$ values for iterations. It is very difficult for both methods to choose the optimal cutoff for feature selection. Fourth, to overcome the aforementioned shortcomings in optimal cutoff determinations for feature selection, the concept of consensus sample classes was introduced in SamCluster system. The identified putative sample classes are not sensitive to the related parameters such as $CV$ thresholds and $P$ values. But for the CLIFF, the number of selected features is chosen empirically during both unsupervised and supervised modes, and the clustering result is sensitive to different choices of this number. Fifth, for the same dataset LEUKEMIA78, our integrated scheme for cluster analysis outperformed the Xing’s method. Moreover, because the gene sets used for the best putative sample class discovery were selected by $t$-test during iteration, the expression level of every gene has a significant difference in two identified putative sample classes.

Based on above results and analysis, we conclude that the integrated scheme combining gene selection based on $CV$ calculation and $t$-test, and hierarchical clustering analysis is an effective method for automatic discovery of sample classes using gene expression profile.
ACKNOWLEDGEMENTS

The work is supported by Beijing 248 Great Creative Project 9550213300 and 01MA078.

REFERENCES


