Interpreting Patterns and Analysis of Acute Leukemia Gene Expression Data by Multivariate Statistical Analysis

ChangKyoo Yoo* and Peter A. Vanrolleghem
BIOMATH, Department of Applied Mathematics, Biometrics and Process Control,
Ghent University, Coupure Links 653, B-9000 Gent, Belgium
(*ChangKyoo.Yoo@biomath.UGent.be)

Abstract

DNA microarray technologies are leading to an explosion in available gene expression data which simultaneously monitor the expression pattern of thousands of genes. Gene expression data are characterized by a very high dimensionality (genes), a relatively small number of samples (observations), irrelevant features, and it leads to a collinearity and multivariate problem. In this paper, we propose a systematic approach to gene selection based on discriminant partial least squares (DPLS) and fuzzy clustering methods. The proposed method was applied to microarray data from leukemia patients; specifically, it was used to interpret the gene expression pattern and analyze the leukemia subtype whose expression profiles correlated with four cases of acute leukemia gene expression.

Keywords: Bioinformatics, classification and clustering, discriminant partial least squares, DNA microarray, gene expression data analysis

1. Introduction

The use of relatively new DNA microarray technologies, which simultaneously monitor the expression pattern of thousands of genes, has led to an explosion of readily available gene expression data. Correspondingly, there now is a great need for methods capable of interpreting, visualizing and analyzing the patterns and information contained within these large data sets. However, gene expression data are characterized by a high dimensionality (genes), a relatively small numbers of samples (observations), irrelevant features, and leads to a collinearity and multivariate problems. Thus, comprehensible interpretation and analysis is difficult and the complexity of the original data entails a high computational cost. To solve the above mentioned problems, the first step in creating such a new method is to extract the fundamental features (or genes) of the gene expression data set (i.e. a dimensional reduction), and the second step is to compare the expression data with the desired level of data analysis (i.e. clustering similar genes or samples, and/or identifying the tumor class). Several studies have used microarray technology to analyze gene expression in colon, breast, leukemia and other cancers (Alizadeh et al., 2000; Cho *et al.*, 2002; Dudoit et al., 2002; Nguyen and Rocke, 2002; Stephanopoulos et al., 2002).

In this paper, we propose a systematic approach to gene selection that uses the discriminant partial least squares (DPLS) and fuzzy clustering methods to interpret the

gene expression patterns of acute leukemia, to identify obscure leukemia subtypes in microarray data, and to establish the relationship between an expression-based leukemia subclass and a clinical outcome.

2. Theory

2.1 Discriminant partial least squares (DPLS)

DPLS is a dimensionality reduction technique for maximizing the covariance between the predictor (independent) block *X* and the predicted (dependent) block *Y* for each component. DPLS models the relationship between *X* and *Y* using a series of local least-squares fits. PLS components are obtained in such a way that the sample covariance between the response variables (leukemia classes) and a linear combination of the predictors (genes), are maximized. In other words, the PLS finds a weight vector **w** which satisfies (Yeung and Ruzzo, 2001; Cho *et al.*, 2002; Nguyen and Rocke, 2002),

$$\mathbf{w_k} = \arg\max \mathrm{Cov}(\mathbf{Xw}, \mathbf{y}) \tag{1}$$

subject to the unit weight and orthogonality constraint

$$\mathbf{w'Sw}_j = 0 \text{ for all } 1 \le j \le k \tag{2}$$

where $\mathbf{S'} = \mathbf{X'X}$. The *i*-th PLS component is a linear combination of the original predictors ($\mathbf{Xw}_i$). The variable importance in the projection (VIP) is a good measure of the influence of all variables in the PLS model on the response variables. The VIP can be calculated from the weight vector of the DPLS model and the percentage that is explained by the dimension of the model, which is defined as follows:

$$VIP = \sum_{a} (\mathbf{w}_{ak})^2 \tag{3}$$

Note that after the PLS weight vectors are computed, genes are selected via the VIP. For a given PLS dimension, $(\mathrm{VIP}_{ak})^2$ is equal to the squared PLS weight $(\mathbf{w}_{ak})^2$. The VIP can be considered as a measure of how much a certain gene corresponds to the samples. Thus, we can select important genes based on the VIP value. It is reasonable to assume that the weights of the features are proportional to their importance in the determination of the class labels; that is, the higher the weight, the better the distinction power of the feature with respect to the class label. Therefore, given a trained PLS classifier, a set of *K* high-ranking genes are obtained by selecting the genes with the top *K* VIP weights.

2.2 Fuzzy c-means (FCM) clustering

In the FCM clustering method, an object can simultaneously be a member of multiple classes (Duda *et al.*, 2001, Yoo *et al.*, 2003). The objective function, which is minimized iteratively, is a weighted within-groups sum of distances $d_{k,i}$. The weighting is performed by multiplying the squared distances by membership values $u_{k,i}$.

$$J_m(C,m) = \sum_{i=1}^{C} \sum_{k=1}^{N} (u_{k,i})^m d_{k,i}^2 \tag{4}$$

where *C* is the total number of clusters, *N* is the total number of objects in the calibration data, $d_{k,i}$ is the distance between an object *k* and a prototype (cluster) *i*, and $u_{k,i}$ is the membership function. After computing the membership values for all calibration objects, the cluster centers (v_i) are described by prototypes, which are fuzzy weighted means, according to the following equation:

$$v_i = \frac{\sum_{k=1}^{N} (u_{k,i})^m x_k}{\sum_{k=1}^{N} (u_{k,i})^m}, \quad \forall i \tag{5}$$

In the prediction of a new test sample, a new value is computed using Eq. (6) (Yoo *et al.*, 2003).

$$u_{N+1,i} = 1 \Big/ \sum_{j=1}^{C} \left(\frac{d_{k,i}^2}{d_{k,j}^2} \right)^{\frac{2}{(m-1)}} \tag{6}$$

For microarray data, we apply the FCM algorithm to the reduced PCA feature space, that is, to the score vector of the PCA. This fuzzy clustering method allows intermediate logical assignments whereby genes or patients are placed into multiple groups by assigning a membership value for each group that is compared between 0 (not in group) and 1 (completely in group). The use of membership values has the advantage of allowing a gene or sample to belong to multiple clusters, which may better reflect the underlying biology.

3. Result and Discussion

The gene selection method proposed in this paper is applied to the acute leukemia data set published by Golub *et al.* (1999). The data set of 7129 genes was derived from 72 patients, 47 of whom were affected with acute lymphoblastic leukemia (ALL; 38 B-cell and 9 T-cell) and 25 of whom were affected with acute myeloid leukemia (AML). The training data set consisted of 38 bone marrow samples, 27 of which were taken from ALL patients (19 B-ALL and 8 T-ALL) and 11 of which were taken from AML patients; the independent (test) data set consisted of 34 patients, 20 of which were taken from ALL patients and 14 of which were taken from AML patients). To remove systematic sources of variation in the microarray, we log transformed the gene expressions, divided the maximum value of each gene and standardized them to have a mean of zero and a standard deviation of one across samples (Yang *et al.*, 2002).

3.1 Interpreting patterns of ALL and AML using FCM

To determine the specific genes that discriminate between ALL and AML, we used the DPLS method for gene selection, where the response variable Y is 0 (AML) or 1 (ALL). Among the 7129 genes, 50 were selected on the basis of the VIP value of DPLS. Thus, on the basis of the correlation coefficients, we chose the 50 genes that were most correlated with the classification of leukemia. In contrast to the 50 genes of Golub *et al.* (1999), the proposed gene selection method assigns high rankings to Zyxin, Leukotriene (C4 synthase gene), Leptin, CD33 antigen, FAH, and Myeloperoxidase (MPO) as well as Cystatins and Cathepsins. These genes are known to play important roles in acute leukemia (Golub *et al*, 1999; Cho *et al.*, 2002). PCA was applied to interpret the patterns of ALL and AML in the leukemia data set because the presence of too many features degrades the clustering performance. The four PCs capture about 77.3% of the variation in the 50 genes by projecting the 50 genes into four dimensions. We applied the FCM clustering method (with four PCs) and analyzed the results of the corresponding clustering and classification. In FCM, the fuzzifier m was set to 1.2 on the basis of the results of many simulations under various conditions. We initialized the

parameters of the cluster prototype center using k-means clustering. Fig. 1 show the FCM membership values for the 38 training samples (left) and the prediction results for the 34 test samples (right). In the training dataset, patients 1-27 have high membership values in class 1 (AML) and low membership values in class 2 (ALL), whereas patients 28-38 show the opposite behavior. Thus, the ALL and AML patients are well clustered without any clustering error. All but two of the 34 test samples were correctly classified. The two misclassified samples were ALL(#42), which showed high gene expression levels in comparison to other ALL patients, and ALL(#66), which showed low expression levels in comparison to other AML patients. This result is superior to that of Golub *et al.* (1999), who obtained a strong prediction for 18 out of the 20 ALL test samples and 10 out of the 14 AML samples (i.e., a total of six misclassifications).

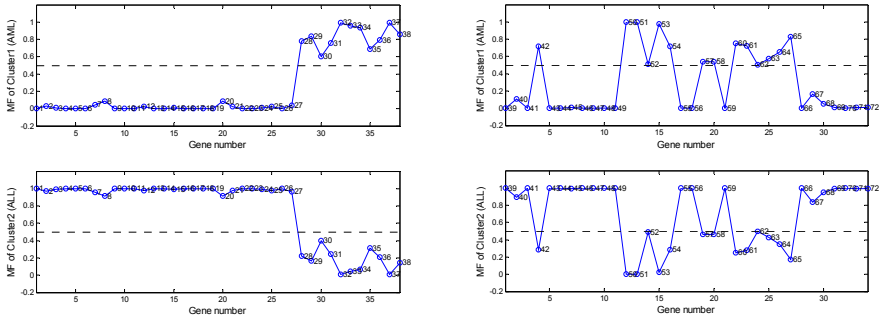

Figure 1. Prediction result of membership values of FCM for training (left) and test samples (right) with cluster 1 (AML, upper) and cluster 2 (ALL, lower).

3.2 Analysis of ALL subclass (B-cell and T-cell)

ALL can be further classified into the T-cell and B-cell lineages. In clinical practice, the B-cell lineage responds better to treatment than the T-cell lineage. Therefore, it is important distinguish between these lineages. To determine the 25 genes that discriminate between T-cell ALL (T-ALL) and B-cell ALL (B-ALL), we used the DPLS method to select the top 25 gene selection, where the response variable Y is 0 (T-ALL) or 1 (B-ALL). After selecting the top 25 genes that are differentially expressed between the B-cell and T-cell lineages of ALL patients, PCA was used to reduce the data dimension. Four PCs were determined, and captured about 83% of the variation in the 25 genes. The result shows that all of the B-cell and T-cell lineage ALL samples are well clustered except for one misclustered sample (#17).

3.3 Analysis of AML subclass: M1, M2, M4, M5

Among the 25 AML patients, we used 20 patients as a training data set, where 4 patients (samples 32, 35, 38, 61) were M1, 10 patients (samples 28, 29, 33, 34, 37, 51, 53, 57, 58, and 60) were M2, 4 patients (samples 31, 50, 52, and 54) were M4, and 2 patients (samples 30 and 36) were M5. The remaining 5 patients (samples 62-66), which could not be classified by Golub *et al.* (1999), were used as a test data set. To determine the genes that discriminate between the AML subclasses included in the training data set (i.e., M1, M2, M4, and M5), we used the DPLS method to select the top 25 gene selection, where the response matrices (Y) were $[1\ 0\ 0\ 0]^T$ for M1, $[0\ 1\ 0\ 0]^T$ for M2, [0

$0\ 1\ 0]^T$ for M4 and $[0\ 0\ 0\ 1]^T$ for M5. After selecting the top 25 genes, PCA was used to reduce the data dimension. Four PCs were determined, and captured about 82% of the variation in the 25 genes. Based on the results of the training data, we used the FCM clustering method to predict the subtype of the five AML samples that could not be predicted by the method of Golub *et al.* (1999). The prediction results indicate that three AML patients (samples #63, 64, and 65) are of subtype M1, and two patients (samples #62 and 66) are of subtype M2.

3.4 Prediction of clinical outcome of AML patients (Failure and Success)

To search for additional sets of genes useful for predicting the clinical outcome of leukemia patients, we performed additional gene selection for the prediction of clinical output of AML treatment. Among the 25 AML patients, we used 15 patients as a training data set, of whom 7 patients (#34-38 and 52-53) survived and 8 patients (#28-33, 50, and 51) died during treatment, and we used the remaining 10 patients (samples # 54, 57, 58, and 60-66), who did not respond to treatment, as a test data set. We used the DPLS method for selecting the top 25 genes, where the response variable Y is 0 (success) and 1 (failure).

After selecting the top 25 genes differentially expressed between AML patients who lived or died during treatment, four PCs using PCA were determined, and captured about 76% of the variation in the 25 genes. In the loading plot (not shown in this paper), genes that correlate with successful treatment appear on the right side and genes that correlate with treatment failure appear on the left side. Almost all of the genes in each gene group have common expression patterns, that is, group-specific regulation patterns known as coregulation patterns (Stephanopoulos et al., 2002). It means that the expression of each group is highly elevated only in the sample class and down-regulated in the other classes. This result is notable in that these genes may be considered marker genes related to the clinical outcome of AML patients.

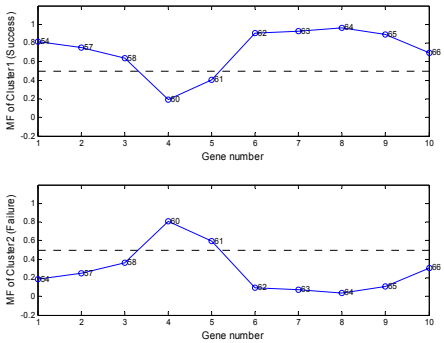

Figure 2. Prediction result of membership values of FCM for 10 test samples (54, 57, 58, 60-66) with AML patients who lived and died after treatment.

Fig. 2 depicts the prediction results based on the membership values from FCM clustering for the 10 AML patients (54, 57, 58, 60-66) whose clinical outcome was not specified by Golub *et al.* (1999). The results indicate that eight AML patients (#54, 57, 58, 62, 63, 64, 65, and 66) are predicted to survive after treatment, and two AML patients (#61 and 62) are predicted to die after treatment. Thus, the proposed method

makes it possible to predict the clinical outcome of AML patients. Moreover, based on the present findings in regard to the link between certain genes and clinical outcome, we can determine the specific genes and relapse in leukemia patients. Although the clinical outcome is also affected by many other factors, such as patient age, treatment regime, and time of diagnosis, the results presented here highlight the potential of the proposed method for uncovering prognostic indicators for leukemia.

4. Conclusion

The DNA microarray technology is useful for discriminating between various subtypes of leukemia, which is necessary for the accurate diagnosis and treatment of patients. Here, we present a novel class-oriented gene selection method and fuzzy clustering method. The proposed method allows the identification of important genes and the classification of leukemia subtype solely on the basis of molecular-level monitoring. Further, the proposed method was used to establish a relationship between expression-based subclasses of leukemia tumors and patient outcome.

5. Acknowledgments

This work was financially supported by the Visiting Postdoctoral Fellowship of the Scientific Research-Flanders (FWO).

6. References

Alizadeh, A., Eisen, and Staudt., L. M. 2000. Different types of diffuse large b-cell lymphoma identified by gene expression profiling, Nature, 403, 503-511.

Cho, J.-H., Lee, D. K., Park, J. H., Kim, K. W. and Lee, I.-B. 2002. Optimal approaches for classification of acute leukemia subtypes based on gene expression data, Biotech. Prog., 18(4), 847-854.

Dudoit, S., Fridlyand, J. and Speed, T. 2002. Comparison of discrimination methods for the classification of tumor using gene expression data, J. Am. Stat. Assoc. 97, 77-87.

Golub T. R., Slonim, D. K., Tamayo, P. and Lander, E. S. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science, 286, 531-537.

Nguyen, D. V. and Rocke, D. M. 2002. Tumor classification by partial least squares using microarray gene expression, Bioinformatics, 18(1), 39-50.

Stephanopoulos G., Hwang, D. H., Schmit, W. A., Misra, J. and Stephanopoulos, G. 2002. Mapping physiological states from microarray expression measurements, Bioinformatics, 18(8), 1054-1063.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P., 2002. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. 30, 15-25.

Yeung, K. Y. and Ruzzo, W. L. 2001. Principal component analysis for clustering gene expression, Bioinformatics, 17, 763-774.

Yoo, C. K., Vanrolleghem, P. A. and Lee, I. 2003. Nonlinear modeling and adaptive monitoring with fuzzy and multivariate statistical method in biological wastewater treatment plant. Journal of Biotechnology, 105(1-2), 135-163.