Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computers and Chemical Engineering 29 (2005) 1345–1356
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

Interpreting patterns and analysis of acute leukemia gene expression data by multivariate fuzzy statistical analysis

ChangKyoo Yoo [a,b,*], In-Beum Lee [b,2], Peter A. Vanrolleghem [a,1]

[a] *BIOMATH, Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, B-9000 Gent, Belgium*
[b] *School of Environmental Science and Engineering, Department of Chemical Engineering, Pohang University of Science and Technology, San 31 Hyoja Dong, Pohang 790-784, South Korea*

Available online 11 March 2005

Abstract

DNA microarray technologies, which monitor simultaneously, the expression pattern of thousands of individual genes in different biological systems have resulted in a tremendous increase of the amount of available gene expression data and have provided new insights into gene expression during development, within disease processes, and across species. However, microarray gene expression data are characterized by very high dimensionality (genes), relatively small numbers of samples (observations), irrelevant features, as well as collinear and multivariate characteristics. These features complicate the interpretation and analysis of microarray data, and the complexity of such data means that its analysis entails a high computational cost. This situation motivated the researchers to develop a new method for analyzing microarray data. In this paper, we propose a simple gene selection and multivariate fuzzy statistical analysis 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; Fuzzy clustering; Gene expression analysis; Gene selection; Molecular biology; Leukemia gene expression; Partial least squares (PLS)

1. Introduction

The recent development of DNA microarray technology, which offers the opportunity to simultaneously study the expression of thousands of individual genes in different biological systems, has provided new insights into gene expression during development, within disease processes, and across species. Recently, researchers have eschewed morphological tumor classifications in favor of classification using gene expression profiles on DNA chips. Therefore, researchers are currently seeking to develop new approaches to (i) diagnose cancer early in its clinical course, (ii) more effectively treat advanced stage disease, (iii) better predict a tumor's response to therapy prior to the actual treatment, and (iv) ultimately prevent disease from arising through chemopreventive strategies. These goals can only be accomplished through a better understanding of how certain genes and their encoded proteins contribute to disease onset and tumor progression, and how they influence the response of patients to drug therapies. Innovations in genetic, biological, biochemical, and data analysis approaches are needed for researchers to fully realize these goals (Ochs & Godwin, 2003).

However, microarray gene expression data are characterized by very high dimensionality (genes), relatively small numbers of samples (observations), irrelevant features, as well as collinear and multivariate characteristics. These features complicate the interpretation and analysis of microarray data, and the complexity of such data means that its analysis entails a high computational cost. In particular, conventional statistical techniques for analyzing gene expression data do not work well (or even at all) when the number of genes far exceeds the number of samples. This situation prompted us

* Corresponding author. Tel.: +82 54 279 5966/+32 9 264 5935; fax: +82 54 279 8299/+32 9 264 6220.
E-mail addresses: ckyoo@postech.edu, ChangKyoo.Yoo@biomath.ugent.be (C. Yoo).
[1] Tel.: +32 9 264 5935; fax: +32 9 264 6220.
[2] Tel.: +82 54 279 5966; fax: +82 54 279 8299.

0098-1354/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2005.02.031

to develop a new method for analyzing microarray data (Lu & Han, 2003).

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).

A lot of studies have used microarray technology to analyze gene expression in colon, breast, leukemia, and other cancers. These studies have demonstrated the ability of expression profiling to cluster similar genes and classify tumors. Gene expression profiles may give more information than traditional morphology. Golub, Slonim, Tamayo, and Lander (1999) used a weighted voting scheme for molecular classification of acute leukemia; this scheme predicts leukemia subtypes by means of a supervised learning algorithm and discriminant decision rules derived on the basis of the magnitude and threshold of the prediction strength. Alon et al. (1999) used a clustering technique based on a deterministic-annealing algorithm to classify cancer and normal colon tissues. Scherf et al. (2000) used average linkage clustering to distinguish between tumor tissues originating from various sites for tumor tissues originating from various site. Alizadeh, Eisen, and Staudt (2000) studied gene expression in the three most prevalent adult lymphoid malignancies. Based on gene expression data, they identified two previously unrecognized types of diffuse large B-cell lymphoma that exhibited distinct clinical behavior. Average linkage hierarchical clustering was used to identify the two tumor subclasses as well as to group genes with similar expression patterns across the different samples. Ross et al. (2000) used cDNA microarrays to study gene expression in the 60 cell lines from the anti-cancer drug screen (NCI 60) of the National Cancer Institute (NCI). Hierarchical clustering of the tumoral cell lines based on gene expression data revealed a correspondence between gene expression and tissue of origin. Hierarchical clustering was also used to group genes with similar expression patterns across the cell lines. Dudoit, Fridlyand, and Speed (2002) compared the result of applying various classifies (such as linear discriminant analysis and quadratic analysis) to the same set of gene expression data and Bicciato, Pandin, Didone, and Di Bello (2002) applied an auto-associative neural network model to pattern identification and classification in gene expression data.

In this paper, we developed a simple 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. Using the DPLS-based gene selection method, we determined the groups of genes whose expression profiles correlated with five cases: (1) acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), (2) ALL subtype (T-cell or B-cell), (3) AML subtype (M1, M2, M4, or M5), and (4) AML subtype by clinical outcome (success or failure). Then, using fuzzy clustering, we could predict the type and subtype of leukemia, identify obscure leukemia subtypes in microarray data, and establish the relationship between expression-based leukemia subclass and clinical outcome.

2. Theory

2.1. Discriminant partial least squares

Nguyen and Rocke (2002) suggested an approach in which high-dimensional vectors are reduced using the partial least squares (PLS) method and then classified using logistic discrimination and quadratic discriminant analysis. They showed that the weight vector of PLS alone could be a good indicator of the correlation between the predictor and response. Cho, Lee, Park, Kim, and Lee (2002) proposed an approach for the construction of the optimal linear classifier based on the genes expression data with PLS. On the other hand, Park, Tian, and Kohane (2002) used the PLS and generalized linear regression methods to link gene expression data with patient survival times and reformulate survival data for a Poisson regression. However, it is more physically reasonable to use all the weight vectors of PLS together with the fraction that is explained by the latent variables.

Discriminant partial least squares (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 $\mathbf{w}$ which satisfies (Nguyen & Rocke, 2002; Yeung & Ruzzo, 2001),

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

subject to the unit weight and orthogonality constraint

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

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. 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:

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

Note that after the PLS weight vectors are computed, genes are selected via the VIP. For a given PLS dimension (VIP_{ak}) 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 clustering

In the fuzzy c-means (FCM) clustering method, an object can simultaneously be a member of multiple classes (Duda et al., 2001; Yoo, Vanrolleghem, & Lee, 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 the total number of objects in the calibration data, $d_{k,i}$ 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).

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

2.3. Interpretation and multivariate fuzzy analysis of gene expression data

Because raw microarray data frequently contain correlations between measured variables and are of high dimensionality, it is necessary to introduce multivariate statistical latent variable methods to increase variable independency and to reduce dimensionality. PCA and PLS are usually used either to reduce the data dimension while retaining the important information or to display the data information in a form that can be easily interpreted. Data clustering can then be applied to the transformed data of lower dimension instead of to the original data. Moreover, the compression of data before clustering causes the clustering algorithm to become more stable and efficient in cases where the original variables are highly correlated and of high dimension (Nguyen & Rocke, 2002).

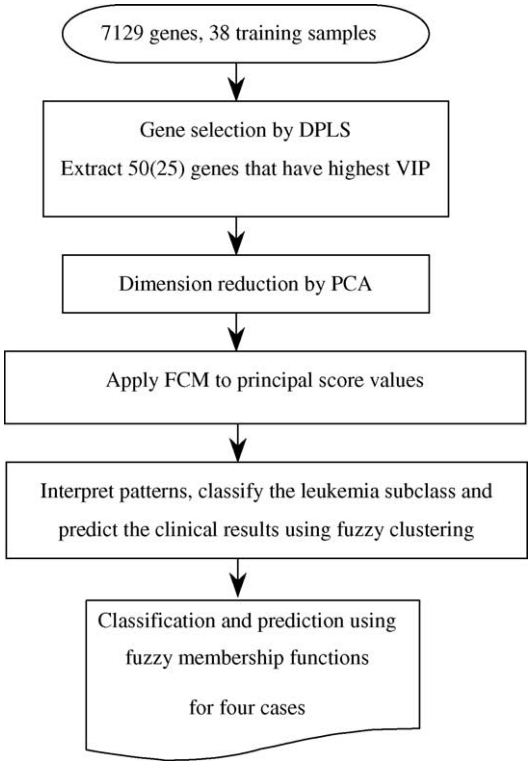

Fig. 1. Schematic flow diagram of the proposed data analysis algorithm for gene expression data.

For microarray data, we apply the FCM algorithm to the reduced PCA feature space, that is, to the score vector of 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 of between 0 (not in group) and 1 (completely in group). The use of membership values has the advantage of allowing gene or sample the possibility of belonging to multiple clusters, which may better reflect the underlying biology.

Fig. 1 shows a schematic flow diagram of the proposed data analysis algorithm for gene selection, dimensional reduction, and the FCM method. First, the relevant genes are selected using the VIP of the DPLS model. Second, the feature dimension is reduced by PCA in order to apply the proposed clustering and prediction method using FCM clustering with Mahalanobis distances for the underdetermined system, i.e., the number of genes is much greater than the number of samples. Third, FCM is used to interpret microarray data patterns, classify the leukemia class, and predict the clinical results of a test sample.

3. Leukemia gene expression microarray dataset

3.1. Background of leukemia

Leukemia is a malignant cancer that originates in cells in the bone marrow, and is characterized by uncontrolled

growth of developing white blood cells. The bone marrow generates cells called blasts that develop (mature) into several different types of blood cells with specific tasks in the human body (Golub et al., 1999): red blood cells (which carry oxygen to all tissues of the body), white blood cells (which fight infection), and platelets (which make the blood clot). There are different types of white blood cells: (i) granulocytes, white blood cells that develop from blood-forming cells called myeloblasts, and mainly destroy bacteria when mature; (ii) monocytes, developing from monoblasts, are also important in protecting the body against bacteria; (iii) lymphocytes are the main cells in lymphoid tissue, which is a major part of the immune system. Two types of lymphocytes are known: B lymphocytes (B-cells) produce antibodies, and T lymphocytes (T-cells) recognize cells infected by viruses and destroy them.

There are two major leukemia classes, myelogenous (also called myeloid) and lymphocytic (also called lymphoblastic) leukemia, which could both be acute or chronic. The terms myelogenous and lymphocytic denote the type of bone marrow cells that is involved. Acute leukemias progress quickly, and can lead to death of a patient within months when not treated. Medical treatment of patients will vary depending on the leukemia class. Thus, knowledge of the leukemia class is very important information for doctors to correctly treat patients. Acute leukemia data sets can be classified into acute lymphoblastic leukemia and acute myeloid leukemia. Moreover, ALL cases can be classified into T-cell ALL and B-cell ALL, depending on the type of lymphocytes that is affected (Golub et al., 1999).

3.2. Gene expression profiles: leukemia dataset

The leukemia dataset and all details with respect to the methods used to collect the data are described in the paper of Golub et al. (1999). The dataset, available at http://www.genome.wi.mit.edu/MPR, consists of a set of high-density oligonucleotide microarrays (Affymetrix) with probes of 7129 human genes, was obtained from 72 patients. Forty seven patients were affected with ALL (38 B-ALL and 9 T-ALL), and 25 patients were affected with AML. The training data set consists of 38 bone marrow samples: 27 samples were taken from ALL patients (19 B-ALL and 8 T-ALL) and 11 were taken from AML patients. The independent (test) data set consisted of 34 samples: 20 ALL patients and 14 AML patients. Furthermore, a description of cancer subtypes, treatment response, patient gender, and laboratory that performed the analysis is provided with the data. Moreover, the result of the subsequent treatment (success or failure) is provided for a limited number of samples. The gene expression profiles of the original data set are represented as log 10 normalized expression values, such that overall intensities for each chip are equivalent. To remove systematic sources of variation in the microarray experiments (i.e., different labeling efficiencies and scanning properties, print-tip or spatial effects, and different noise levels in each array), the expression level of each gene was normalized to have a zero mean and a standard deviation of one (Yang et al., 2002).

4. Result and discussion

The proposed method is applied to the acute leukemia data set published by Golub et al. (1999) with four cases: (1) acute lymphoblastic leukemia and acute myeloid leukemia, (2) ALL subtype (T-cell or B-cell), (3) AML subtype (M1, M2, M4, or M5), and (4) AML subtype by clinical outcome (success or failure).

4.1. Interpreting patterns of ALL and AML using fuzzy clustering

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, where the VIP plot of DPLS for the leukemia data set is displayed in Fig. 2. Thus, on the basis of the correlation coefficients, we chose the 50 genes that were most correlated with the classification of leukemia. The top 50 genes were examined for chromosomal localization using NCBI LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink). In contrast to the 50 genes of Golub et al. (1999), it 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. For example, Zyxin is located in chromosome 7, which may contain genes related to myeloid malignancy, and Cystatins are endogenous protein inhibitors of cathepsins, and hence these specific protease-inhibitors might be important in the etiology of ALL and AML. In addition, CD33 is located in chromosome 19q13.3, and has been developed for targeted antibody therapy to kill leukemia AML cells (Thomas, Olson, Tapscott, & Zhao, 2001). Almost all of the

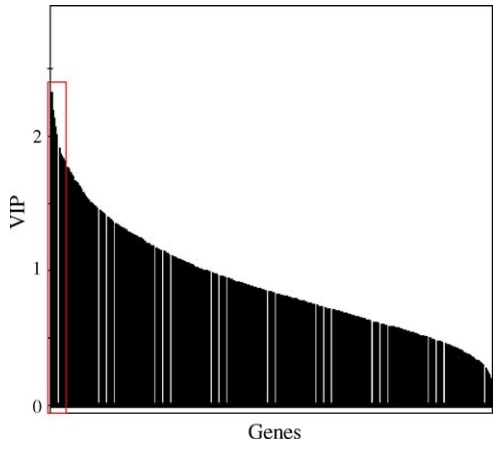

Fig. 2. Variable influence on projection (VIP) of DPLS in the leukemia data set. The higher the VIP value, the more influential the variable.

selected genes have a high expression level (more than 10^3) and show large discrepancies between the ALL and AML samples. This means that the genes selected may hardly be contaminated by noise, key discriminating genes between ALL and AML, and hence, they are highly likely to be critical candidates for the distinction of leukemia subtypes.

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 optimal number of PCs should be determined considering both the curse of dimensionality and the loss of information. Several techniques exist for determining the optimal number of PCs, but to date no dominant technique has emerged. In the system considered here, four PCs were found to be adequate based on the cross-validation of the prediction residual sum of squares (PRESS). The four PCs capture about 77.3% of the variation in the 50 genes by projecting the 50 genes into four dimensions. To interpret the leukemia data set, we examined the score plots of the training and test data in the two-dimensional space spanned by the first two PCs, which are shown in Fig. 3. Fig. 3(a) shows the PCA score plot for 38 training data sets of ALL and AML. For the validation, the projected samples of an independent matrix of expression data from 34 patients is shown in Fig. 3(b). Here, ALL and AML can be visually distinguished with the exception being the 66th patient. Thus, dimension reduction by PCA can make it possible to distinguish between ALL and AML.

We now consider why the 66th patient was abnormal. To investigate the source of misclassification of this patient or more specifically to identify which genes are responsible for this particular patient having cancer, we examined the contribution plot and gene expression of the abnormal 66th AML patient, which are illustrated in Fig. 4(a). Contribution plots are graphical representations of the contribution of each gene to the deviation of the current patient from that defined by the PCA model for the learning data set. By interrogating the underlying PCA model at the point where the abnormality is detected, one can extract diagnostic or contribution plots that reveal the group of genes making the greatest contributions to the deviations in the scores. Inspection of such plots could potentially reveal the group of genes that most influence the difference between the ALL and AML patients. The contribution plot derived from the mean values enables classification of the patients as having ALL or AML. The contributions of most genes in the 66th patient are negative, contrary to the behavior of the other AML patients. These plots provide considerable insight into the possible factors causing this patient to appear abnormal in the current analysis, and thus, greatly narrow the search for the source of the abnormality. For a detailed illustration, Fig. 4(b) shows the gene expression levels of the top five genes for patient 66 and patients 50–54. This plot reveals that patient 66 has low gene expression levels, compared to other AML patients. In particular, the top-ranked gene, Zyxin, shows an abnormally low expression level for patient 66.

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. 5 shows 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 2 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.

4.2. Analysis of ALL subclass (B-cell and T-cell)

Acute lymphoblastic leukaemia is a heterogeneous disease with distinct biological and prognostic groupings. Diagnosis relies on traditional cytomorphological and immunohistochemical evaluation of the leukaemic blasts. Subsequently, cytogenetic analysis identifies clonal numeric and/or structural chromosomal abnormalities that may be present, thus confirming the subtype classification and providing important prognostic information for treatment planning (Kebriaei, Anastasi, & Larson, 2002). 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. Among the 47 ALL patients of Golub et al. (1999), 27 patients were used as a training data set (19 B-cell ALL and 8 T-cell ALL). 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).

In general, ALL is cytogenetically classified as belonging to one of the following classes: (i) hyperdiploid (more than 50 chromosomes); (ii) pseudodiploid (abnormal 46 chromosomes); (iii) diploid (normal 46 chromosomes); (iv) hypodiploid (less than 46 chromosomes). Numerous chromosomal translocations have been associated with the disease, some of which occur only rarely. ALL breakpoints often involve the immunoglobulin (lg) (B-ALL) or T-cell receptor (T-ALL) genes. The two most common translocations are t(9;22) and t(11q23). In addition, the t(1:19) translocation is common in childhood B-ALL, and t(8;14)(q24;q32), and variants t(2;8) and t(8;22) are found in almost all B-ALL patients. T-ALL patients may show t(1;14)(p32;q11) or 14q11, 7q34-36 or 7p15 translocations, which involve the T-cell receptor loci. The Philadelphia chromosome [t(9;22)(q34;q11)] is found in about 30% of adults with ALL (Golub et al., 1999; Yeoh et al., 2002). Almost all selected 25 genes mapped to regions that have been previously associated with ALL chromosomal

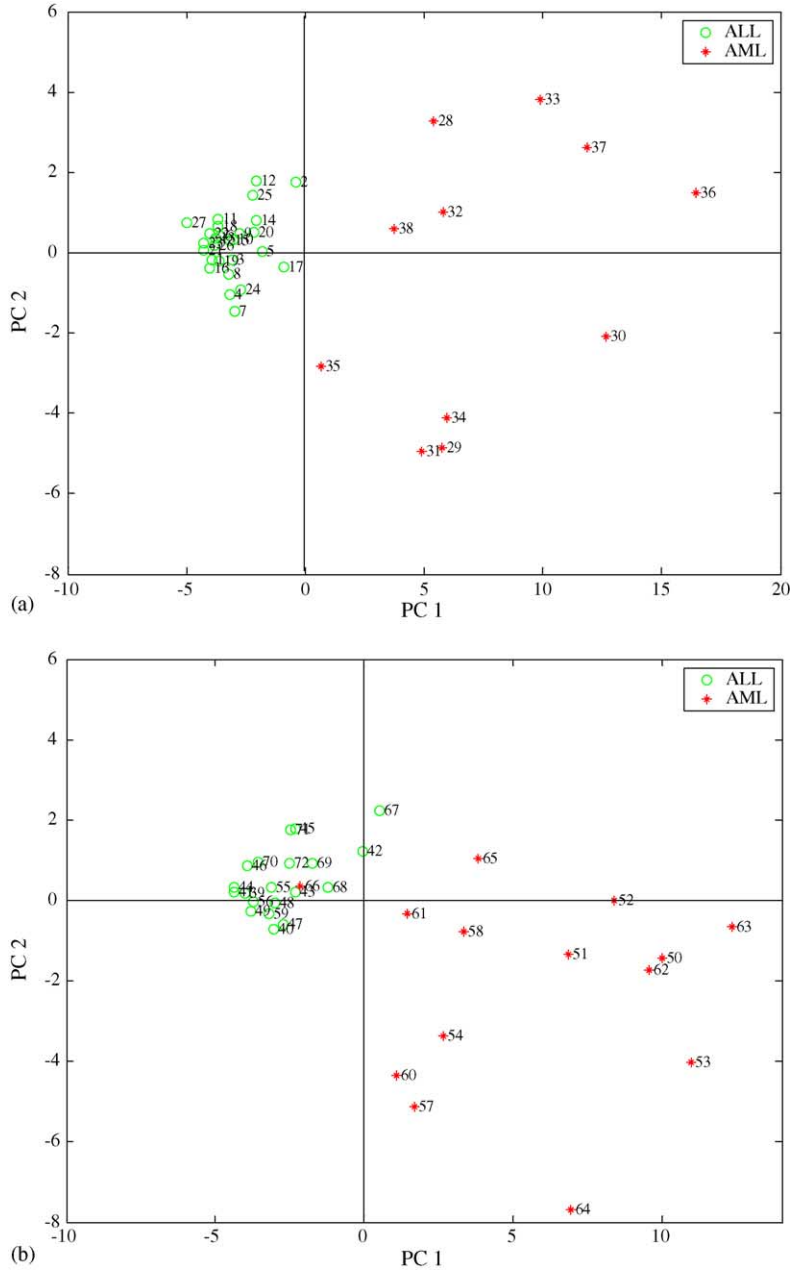

Fig. 3. Score plot of PCA for ALL and AML leukemia: (a) training samples of 38 patients and (b) test samples of 34 patients.

abnormalities, including the T-cell antigen receptor (X03934, 9p56), TCRB (X00437, 7q34) (CD47, X69398 3q13), CD7 (D00749, 7q34), and TCF7(X59871, 5q31). Kebriaei et al. (2002) report that the major chromosomal abnormalities in ALL are t(9;22)(q34;q11), t(12;21)(p13;q22), t(4;11)(q21;q23), t(1;19)(q23;p13), 8q24 translocations and hyperdiploidy. These results suggest that the 25 genes selected via DPLS as being most relevant for classifying B-ALL and T-ALL subclasses are biologically relevant as well.

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. Fig. 6 shows the clustering results with FCM membership values for ALL samples of 47 patients with T-cell and B-cell lineages. All of the B-cell and T-cell lineage ALL samples are well clustered except for one misclustered sample (#17). These results confirm that the top 25 genes are differentially expressed between the T-ALL and B-ALL subclasses of ALL patients.

4.3. Analysis of AML subclass: M1, M2, M4, M5

The original French–American–British (FAB) system for determining leukemia subtype was based only on the appearance of leukemic cells under the microscope after routine

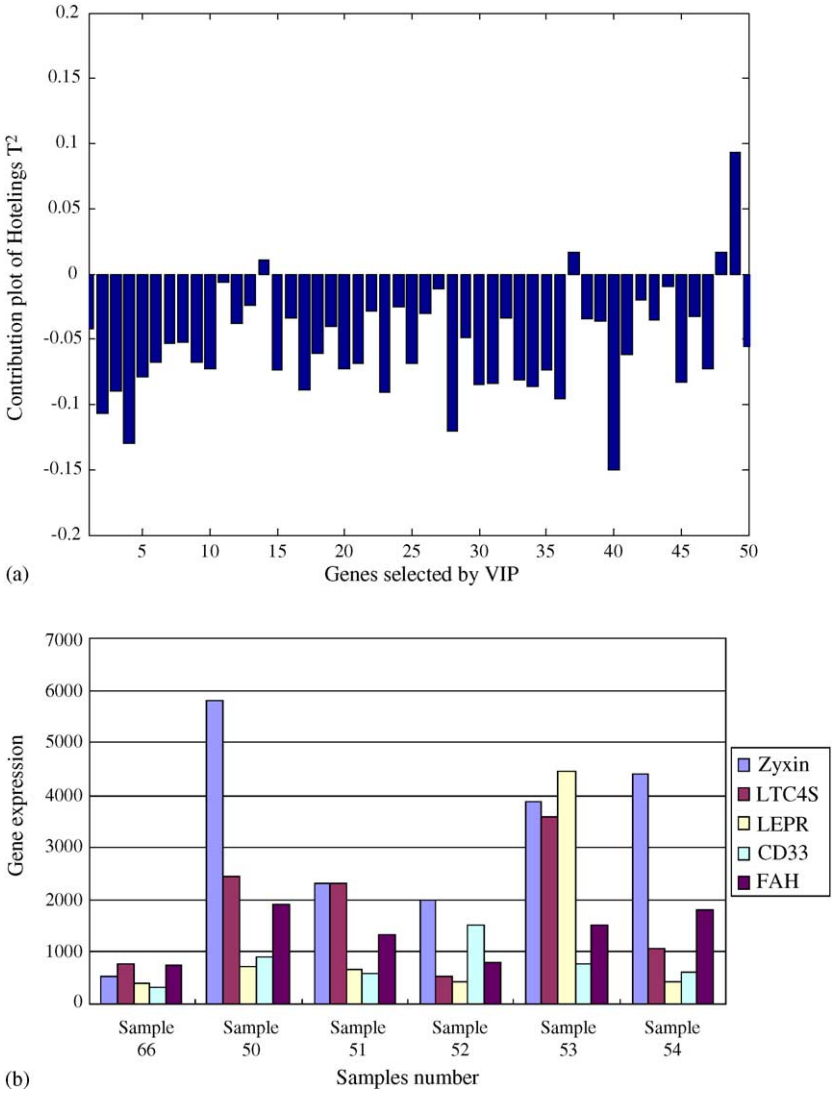

Fig. 4. Illustration for the abnormal 66th AML sample: (a) contribution plot and (b) comparison of gene expression levels of top five genes (samples 66 and 50–54).

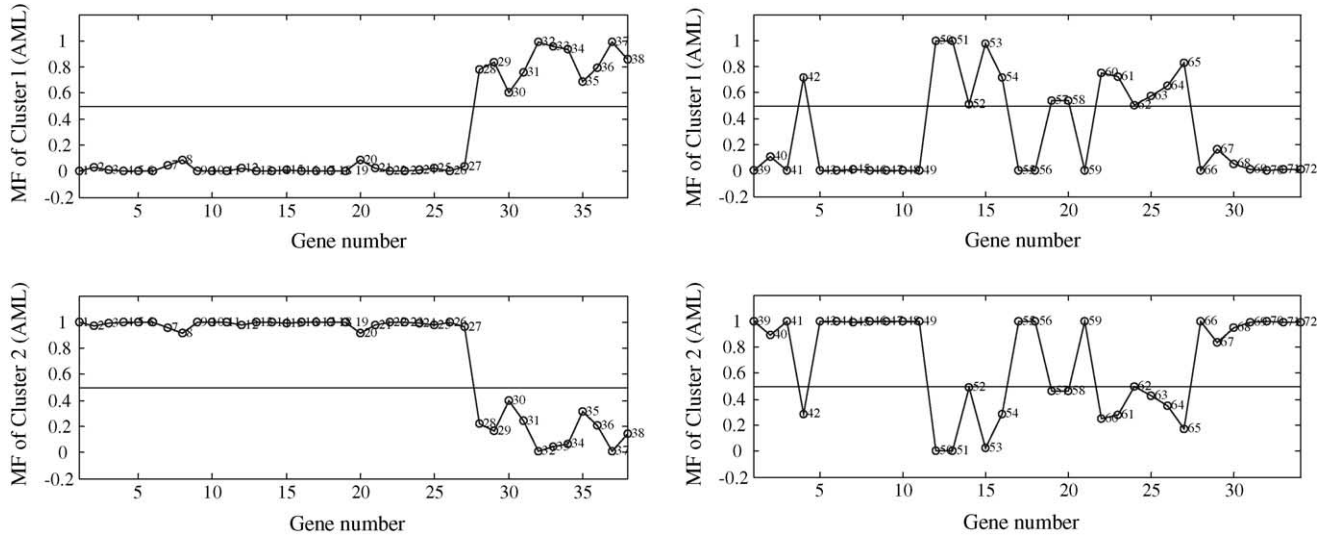

Fig. 5. Prediction result of membership values of FCM for training (left) and test samples (right) with cluster 1 (AML, upper) and cluster 2 (ALL, lower).

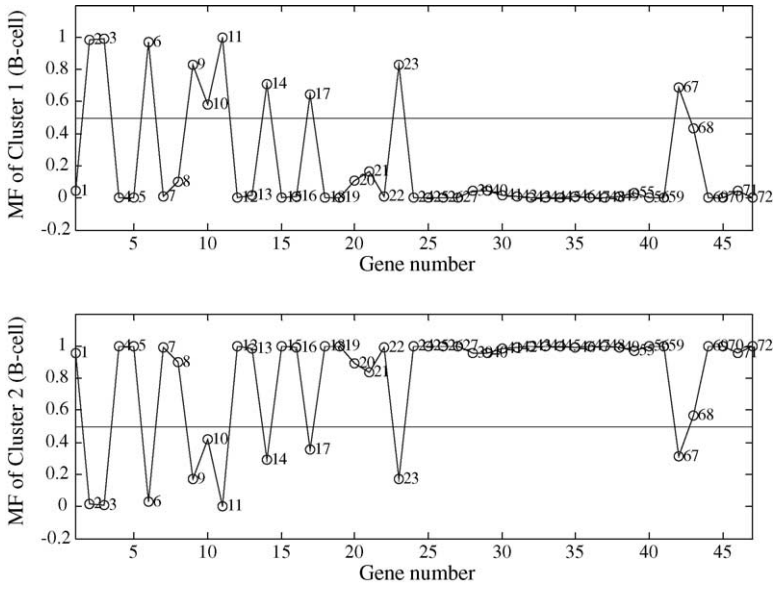

Fig. 6. FCM clustering results of ALL samples of 47 patients: (a) cluster 1 (T-cell lineage) and (b) cluster 2 (B-cell lineage).

processing or cytochemical staining. AML can be classified into six subtypes, designated M1–M6. Although patients tend to be classified into either the M2 or M4 subtype under the FAB system, it is difficult to for most doctors to discriminate sharply between these subtypes. And identifying the M3 subtype is of importance because this subtype usually responds well to treatment with retinoids. The M5 subtype is not easy to detect using the FAB system and usually shows poor response to treatment; most doctors recommend intensive chemotherapy for patients with this subtype. Correct identification of the AML subtype is very important to the clinical treatment step. Because the AML subtype cannot be determined in some patients, leukemia should be assigned to more than one cluster (Golub et al., 1999).

Among the 25 AML patients, we used 20 patients as a training data set, where 4 patients (samples 32, 35, 38, and 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 five 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. We selected the top 25 genes of the AML subclass with M1, M2, M4, and M5. The Philadelphia chromosome is found in less than 1% of AML patients. Other genetic abnormalities associated with AML include t(8;21)(q22;q22), which is observed most frequently in children and young adults and is associated with the M2 subtype. Almost all patients with AML M3 show t(15;17)(q22;q21), which affects the retinoic acid receptor alpha and thus leads to acute promyleocytic leukemia. In addition, t(11q23) occurs frequently in both ALL and AML patients. Many of the genes in the top 25 genes of the AML subclass with M1, M2, M4, and M5 encode proteins critical for S-phase cell cycle progression (Cyclin D3, Op18, and MCM3), chromatin remodeling (RbAp48 and SNF2), transcription (TFIIE β), and cell adhesion (zyxin) or are known oncogenes (c-MYB, E2A and HOXA9). CD33 and MB-1 encode cell surface proteins for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells (Dorrie, Gerauer, Wachter, & Zunino, 2001).

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. Fig. 7 shows the PCA score plot of the 20 AML patients with M1, M2, M4, or M5 comprising the training data set and the five test samples

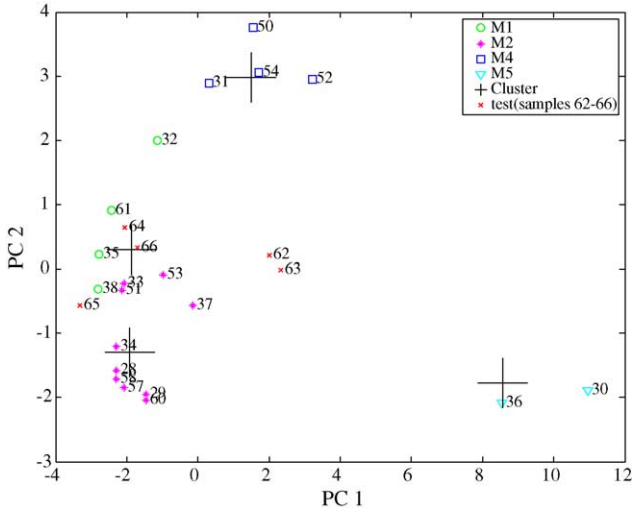

Fig. 7. PCA score plot of 25 AML patients with M1, M2, M4, or M5, and five test samples (62–66).

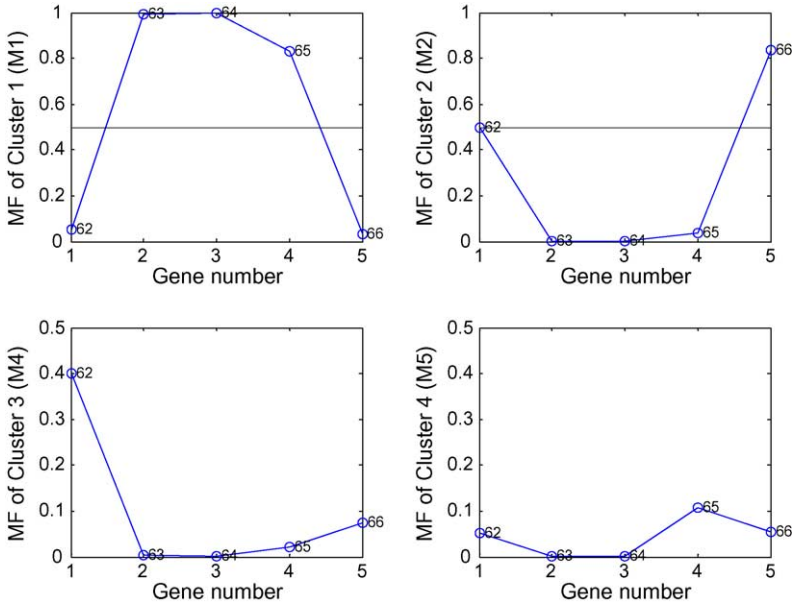

Fig. 8. Prediction result of membership values of FCM for five test AML samples (62–55) with subclass M1, M2, M4, or M5.

(62–66) plotted in the two-dimensional space spanned by the first two PCs. In this plot, the AML patients are well separated into four clusters, one for each subtype, without any clustering error. 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). Fig. 8 depicts the prediction results for the five unknown test samples (#62–66) using FCM clustering. The prediction results indicate that three AML patients (samples #63–65) are of subtype M1, and two patients (samples #62 and 66) are of subtype M2.

4.4. Prediction of clinical outcome of AML patients (failure and success)

The genomewide expression patterns of tumors provide a good representation of their biology and diversity. Thus, relating gene expression patterns to clinical outcomes is a key issue in cancer genetics. Many parameters have been explored in relation to leukemia cancer biology and disease outcome, and researchers have found that some patients is a good indicator of prognosis. However, some patients have been found to opposite prognosis, as well, which underlines the difficulties of correlating single factors with prognoses (Thomas et al., 2001). Although the FAB system for classifying AML as M1, M2, M3, M4, M5, or M6 by morphological states is based on clinical data, we should elucidate the factors underlying the success or failure of treatment, which would allow us to better predict the *clinical outcome* of leukemia patients. One of the most promising aspects of gene expression profiling is the hope that it will enable more accurate identification of patients who are at a high risk of failing conventional therapy.

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 for discriminating between failure and success of clinical treatment of AML patients. The chromosomal locations of the 25 identified genes were checked in the NCBI LocusLink, because chromosomal abnormalities are prevalent in leukemia patients and often have prognostic implications (Thomas et al., 2001). Almost all genes among the selected 25 genes have been identified previously as containing abnormalities in AML or another form of leukemia. Most of the genes reported by Lyons-Weiler, Patel, and Bhattacharya, 2003 are also found in our marker gene set (HoxA9, PIG-B, MACH-alpha-2 protein, BPI Bactericidal/permeability increasing protein, Autoantigen PM-SCL, ERGIC-53 Protein, and so on). Overexpression of HoxA9 would presumably result in an overproduction of leukocytes and lymphocytes. Indeed, the injection of retrovirally engineered primary bone marrow cells that overexpresses both HoxA9 and Meis1 into mice induces AML within three months (Kroon et al., 1998). Golub et al. (1999) found that HoxA9 had the highest correlation to their ideal distribution, but did not find a suitable gene set that enabled predicting chemotherapy success and failure. Thomas et al. (2001) suspected that, out of all the genes in the original data, HoxA9 could predict success and failure of chemotherapy, but were confronted with a lack of statistical significance in their measure of the difference between success and failure ($P < 0.1$). When checking the gene expression profiles of HoxA9 among the 15 AML patients in the training group, those with poor treatment outcomes

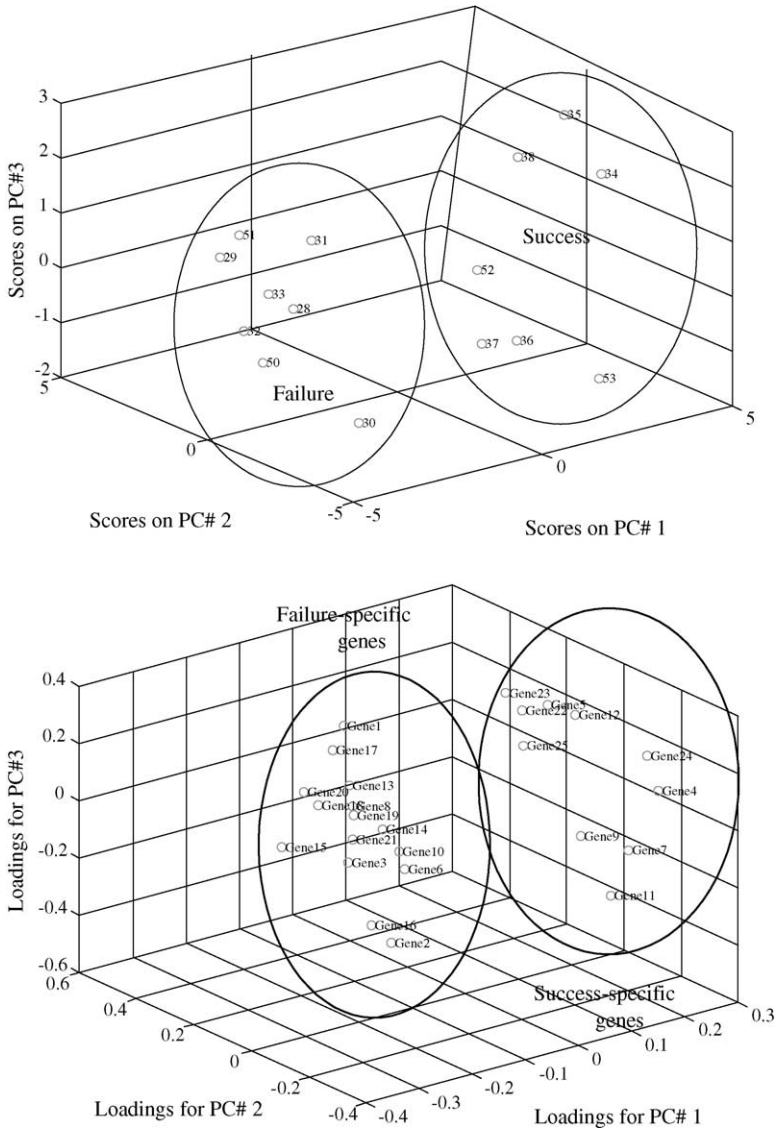

Fig. 9. Analysis and interpretation using PCA for 15 AML patients (8 failure and 7 success samples): (a) score plot and (b) loading plot.

among these patients showed increased expression of HoxA9.

After selecting the top 25 genes, differentially expressed between AML patients who lived or died during treatment, PCA was used to reduce the data dimension. Four PCs were determined, and captured about 76% of the variation in the 25 genes. To determine the correlation between the genes selected for AML patients and clinical outcome, we examined the PCA score and loading plots of the 15 AML patients (8 failure and 7 success samples); these plots are shown in Fig. 9. These plots demonstrate that the selected 25 genes can visually discriminate the clinical outcome of AML patients, and that PCA can extract the key feature components. The plot of the PCA loadings (Fig. 9(b)) can be used to establish how the 25 genes are interrelated. The form of the loading plot is closely connected with the pattern of the score plot (Fig. 9(a)), and shows how the 25 genes are expressed and how they interact to separate the AML patients based on clinical outcome. In the loading plot, 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. It means that the expression of each group is highly elevated only in the sample class and down-regulated in the other classes (Stephanopoulos, Hwang, Schmit, Misra, & Stephanopoulos, 2002). This result is notable in that these genes may be considered marker genes related to the clinical outcome of AML patients.

Fig. 10 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, 62–66) 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

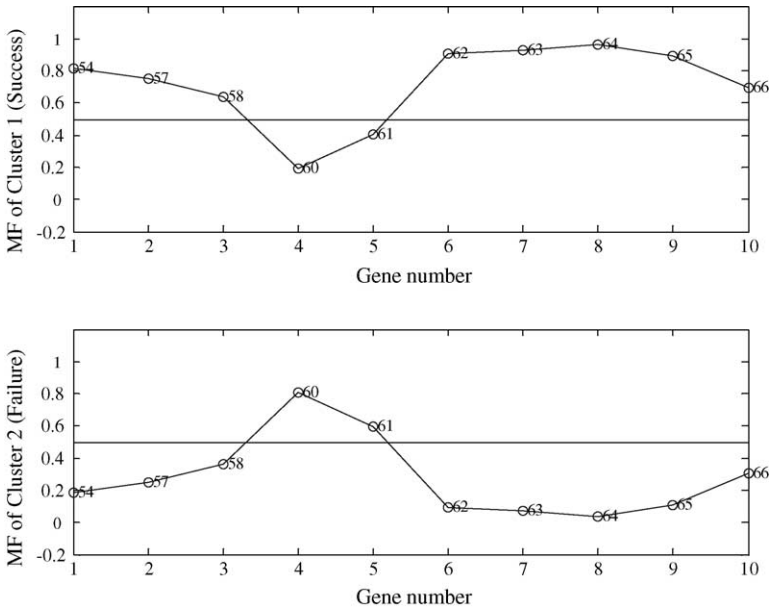

Fig. 10. Prediction result of membership values of FCM for 10 test samples (54, 57, 58, 60–66) with AML patients who lived and died after treatment Fig. 3. Trajectories of nine variables from a nominal batch run.

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, and also suggest a medicine manufacturer to make a new drug development with the maker genes, which manifest themselves among the survival patients after treatment. 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.

5. Conclusions

Biotechnological advances, such as DNA microarray analysis of gene expression, allow researchers to enlarge their understanding of living systems, biochemical pathways, and even disease. Indeed, 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 simple class-oriented gene selection method and fuzzy clustering method. This new method is a simple and efficient way to identify genes and gene expression signatures that may be used to distinguish between leukemia classes and subclasses. Additionally, a fuzzy clustering method was proposed to solve the problems encountered using existing partitional clustering and threshold-based clustering methods, which assign each sample to a single class. The present results demonstrate the utility of fuzzy clustering for elucidating the complex modes of gene expression regulation, for extracting biological insights from microarray data, and for identifying leukemia subtypes. It makes possible the identification of important genes for each subclass and the classification of leukemia subtype solely on the basis of molecular-level monitoring. It was also used to establish a relationship between expression-based subclasses of leukemia tumors and patient outcome, which can give a hint to the drug development. Thus, it can potentially be used to guide the design of new, more effective approaches to the treatment of leukemia. Because the proposed method is based on a simple gene selection and fuzzy clustering methods, it can be used in other microarray data. We are developing a quite innovative method for a simultaneous classification and an unknown subclass finding with a new generic statistic so as to remove the limitations of a threshold-based gene selection, such as an inability of an unknown subclass.

Acknowledgements

This work was supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF) and a Visiting Postdoctoral Fellowship of the Fund for Scientific Research-Flanders (FWO).

References

Alizadeh, A., Eisen, M. B., & Staudt, L. M. (2000). Different types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503–511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America, 96*, 6745–6750.

Bicciato, S., Pandin, M., Didone, G., & Di Bello, C. (2002). Pattern identification and classification in gene expression data using an autoassociative neural network model. *Biotechnology and Bioengineering*, *81*, 594–606.

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

Dorrie, J., Gerauer, H., Wachter, Y., & Zunino, S. J. (2001). Resveratrol induces extensive apoptosis by depolarizing mitochondrial membranes and activating caspase-9 in acute lymphoblastic leukemia cells. *Cancer Research*, *61*, 4731–4739.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). New York: John Wiley & Sons.

Dudoit, S., Fridlyand, J., & Speed, T. (2002). Comparison of discrimination methods for the classification of tumor using gene expression data. *Journal of American State Association*, *97*, 77–87.

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

Kebriaei, P., Anastasi, J., & Larson, R. A. (2002). Acute lymphoblastic leukaemia: diagnosis and classification. *Best Practice and Research. Clinical Haematology*, *15*, 597–621.

Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. *Information Systems*, *28*, 243–268.

Lyons-Weiler, J., Patel, S., & Bhattacharya, S. (2003). A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Research*, *13*, 503–512.

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

Ochs, M. F., & Godwin, A. K. (2003). Microarrays in cancer: Research and applications. *BioTechniques*, *34*, S4–S15.

Park, P. J., Tian, L., & Kohane, I. S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, *18*(1), S120–S127.

Ross, T., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, *24*, 227–234.

Scherf, U., et al. (2000). A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, *24*, 236–244.

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

Stephenson, J. (1999). Human genome studies expected to revolutionize cancer classification. *Journal of American Medical Association*, *282*, 927–992.

Thomas, J. G., Olson, J. M., Tapscott, S. J., & Zhao, L. P. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research*, *11*, 1227–1236.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., & Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, *30*, 15–21.

Yeoh, E., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, *1*, 133–143.

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression. *Bioinformatics*, *17*, 763–774.

Yoo, C. K., Vanrolleghem, P. A., & 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), 135–161.