Journal of Chemical Engineering of Japan, Vol. 41, No. 9, pp. 898–914, 2008 **Research Paper**

Classification and Diagnostic Output Prediction of Cancer Using Gene Expression Profiling and Supervised Machine Learning Algorithms

Changkyoo YOO[1] and Krist V. GERNAEY[2]
[1]*College of Environment and Applied Chemistry, Green Energy Center/Center for Environmental Studies, Kyung Hee University, Seocheon-dong 1, Giheung-gu, Yongin-Si, Gyeonggi-Do, 446-701, Korea*
[2]*Department of Chemical Engineering, Technical University of Denmark, Building 229, DK-2800 Kgs. Lyngby, Denmark*

Keywords: Bioinformatics, Data Mining, Gene Expression Profiling, Cancer Classification, Supervised Clustering

In this paper, a new supervised clustering and classification method is proposed. First, the application of discriminant partial least squares (DPLS) for the selection of a minimum number of key genes is applied on a gene expression microarray data set. Second, supervised hierarchical clustering based on the information of the cancer type is subsequently proposed to find key gene groups and to group the cancer samples into different subclasses. Here, the weights of the genes in the DPLS are proportional to their importance in the determination of the class labels, that is, the variable importance in the projection (VIP) information of the DPLS method. The power of the gene selection method and the proposed supervised hierarchical clustering method is illustrated on a three microarray data sets of leukemia, breast, and colon cancer. Supervised machine learning algorithms thus enable the subtype classification 3 data sets solely on the basis of molecular-level monitoring. Compared to unsupervised clustering, the supervised method performed better for discriminating between cancer types and cancer subtypes for the leukemia data set. The performance of the proposed method, using only a limited set of informative genes, is demonstrated to be comparable or better than results reported in the literature for the three data sets. Furthermore the method was successful in predicting the outcome of medical treatment (success or failure) based on the microarray data, which could make the method an important tool for clinical doctors.

Introduction

The use of the relatively new DNA microarray technology, which enables simultaneous monitoring of the expression pattern of thousands of genes, has led to an explosion in the amount of readily available gene expression data. Correspondingly, there now is a great need for methods capable of interpreting, visualizing and analyzing the gene expression pattern data. However, analyzing gene expression data is far from straightforward (Lu and Han, 2003). The data are typically characterized by a very high dimensionality (high number of genes), a relatively small number of samples (observations), irrelevant features, as well as collinear and multivariate characteristics. In particular, conventional statistical techniques do not work well (or even not at all) for analysis of gene expression data, when the number of variables (genes) by far exceeds the number of samples. Thus, there is great need for new methods that are capable of analyzing microarray data. The first step in creating such a new method consists of extracting the fundamental features (or genes) of the gene expression data set (i.e. a dimensionality reduction). The second step is the usage of the retained expression data within the desired framework of data analysis, which could for example be classifying similar genes or samples, and/or identifying the tumor class for a given sample (Dudoit *et al*., 2002).

A lot of studies have used microarray technology to analyze gene expression in colon, breast, leukemia and other cancers (Golub *et al*., 1999; Alizadeh *et al*., 2000; Furey *et al*., 2000; Quackenbush, 2001; Cho *et al*., 2002; Dudoit *et al*., 2002; Nguyen and Rocke, 2002a, 2002b; Stephanopoulos *et al*., 2002; Hampton and Frierson, 2003; Ishida *et al*., 2003; Lu and Han, 2003; Takahashi *et al*., 2003, 2004, 2005a; Kulkarni *et al*., 2005; Takahashi and Honda, 2005; Bullinger *et al*., 2007). These studies have demonstrated the ability of expression profiling to cluster similar genes and classify tumors. Lu and Han (2003) provide a detailed review on methodologies that are commonly used for microarray gene expression analysis.

Received on March 3, 2008; accepted on May 22, 2008. Correspondence concerning this article should be addressed to C.K. Yoo (E-mail address: ckyoo@khu.ac.kr or ChangKyoo.Yoo@biomath.ugent.be).

Many machine learning methods using support vector machines (SVM) and Fisher's linear discriminant analysis (FLD) on gene expression profiles have recently been applied in cancer classification for colon and breast cancer, for leukemia and other tumors. (Cristianini *et al*., 2000; Moler *et al*., 2000; Furey *et al*., 2000; Xiong *et al*., 2001; Zhang *et al*., 2001; Dudoit *et al*., 2002; Stephanopoulos *et al*., 2002). These investigations have clearly shown the capability of gene expression profiling for classifying the tumors. Gene expression profiles may give more objective information than traditional morphological tumor characterization methods.

On the other hand, there are a number of research works about the methodologies to classify samples into subclasses by the selected genes and the methodology to select gene for such purpose. Bhattacharjee *et al*. (2001) suggested a hierarchical and probabilistic clustering of expression data defined distinct subclasses of lung adenocarcinoma. Two genes of neuroendocrine genes and of type II pneumocyte genes with high relative expression are selected for the tumor subclass. It was revealed a less favorable outcome for the adenocarcinomas with neuroendocrine gene expression. Tibshirani *et al*. (2002) suggested an approach to cancer class prediction from gene expression profiling, based on an enhancement of the simple nearest prototype classifier for subclass cancer. Ishida *et al*. (2003) suggested a new methodology for key gene selection and clinical result for the classification of synthetic retinoids and retinoid synergists. Fifty marker genes whose expression pattern could distinguish these classes are selected by analyzing the effects of all-trans retinoic acid and 9-cis retinoic acid on the gene expressions in a leukemia cell line. And then they found the existence of two subclasses among the selected genes. Bullinger *et al*. (2007) analyzed the AML patients with CBF leukemia using DNA microarray technology and correlated findings with known collaborating aberrations in CBF AML It leads to the identification of clinically relevant subclasses, highlighting genes and pathways of potential pathogenic relevance that provide a basis for novel molecular targeted therapeutic approaches.

In this paper, we first explain the gene selection method, discriminant partial least squares (DPLS), which was used for the selection of the key genes. Secondly, various supervised classification methods, such as principal component analysis (PCA), Fisher's linear discriminant (FLD) analysis and support vector machines (SVM) are subsequently used to classify the gene expression data sets. Third, a new supervised hierarchical clustering method then is proposed using information obtained from the DPLS. The results in microarray dataset shows that that the proposed method allows prediction of tumor type and subtype for three microarray data sets from leukemia, breast and colon cancer patients, as well as establishment of the relationship between expression-based subclass and clinical treatment outcome. Performance of the classification methods is compared with other results reported in the literature.

1. Material and Methods

1.1 Gene selection and dimension reduction in gene expression data

Gene selection (feature selection) is a fundamental issue in gene expression data based tumor clustering and classification. In our research we used discriminant partial least squares (DPLS) as selection method (Nguyen and Rocke, 2002a, 2002b; Sun, 2004a, 2004b). In DPLS, $X \in \Re^{n*m}$ corresponds to a gene expression data and each column in $Y \in \Re^{n*p}$ corresponds to a class. Each element of Y is either 1 or 0. The DPLS method can be explained mathematically as follows: DPLS components are obtained in such a way that the sample covariance between the response variables (in this case the cancer subclasses) and a linear combination of the predictors (genes) are maximized. In other words, DPLS finds a weight vector $\mathbf{w}$ which satisfies (Yeung and Ruzzo, 2001; Cho *et al*., 2002; Nguyen and Rocke, 2002a, 2002b; Sun, 2004a, 2004b; Yoo *et al*., 2005):

$$\mathbf{w}_k = \arg\max \operatorname*{cov}_{w'w=1}^{2}(\mathbf{Xw}, \mathbf{y}) \tag{1}$$

subject to the orthogonality constraint

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

where $\mathbf{S'} = \mathbf{X'X}$. The i-th DPLS component is a linear combination of the original predictors ($\mathbf{Mw}_i$).

In the DPLS method, gene components are selected by sequentially maximizing the covariance between the cancer types and a linear combination of the genes (also subject to orthogonality and normality constraints). This procedure identifies the gene component weights, w, for which cov($\mathbf{X}w$, y) reaches a maximum, where y is the response vector of cancer subtypes. Note that once the DPLS weight vectors are computed, relevant genes are selected via the variable importance in the projection (VIP), which is defined as follows (Eriksson *et al*., 1995):

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

where $\mathbf{w}_{\mathrm{ak}}$ is the PLS weight for the gene expression profiles. The VIP is the sum over all model dimensions of the contributions and can be considered as a good measure of the influence of all genes in the model on the cancer class prediction. For a given DPLS dimension,

VIP_k 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, that is genes that allow to discriminate between different cancer classes, based on the VIP value. This concept is similar to that underlying the selection of genes on the basis of the weights of a linear discriminant function, whereby the genes with the top K weights are selected (where K is the desired number of the selected genes). Therefore, given the DPLS model, a set of K high-ranking genes is obtained by selecting the genes with the top K VIP weights (Sun, 2004a, 2004b; Yoo *et al*., 2005).

In spite of applying gene selection of the original data set, microarray data may still contain redundant information. In this paper, two methods of principal component analysis and Fisher's linear discriminant analysis are used for the dimension reduction after the gene selection. Principal component analysis (PCA) is a dimensionality reduction technique that uses a linear transformation to sequentially maximize the variance of a linear combination of the predictor variables (Nguyen and Rocke, 2002a),

$$\mathbf{v}_k = \underset{\mathbf{v}'\mathbf{v}=1}{\operatorname{argmax}} \operatorname{var}^2(\mathbf{X}\mathbf{v}) \tag{4}$$

subject to the orthogonality constraint

$$\mathbf{v}'\mathbf{S}\mathbf{v}_j = 0, \text{ for all } 1 \le j \le k \tag{5}$$

where $\mathbf{S}' = \mathbf{X}'\mathbf{X}$ is the covariance matrix (Quackenbush, 2001). A lot of research work using PCA and SVD has been performed for analyzing and classifying gene expression data (Alter *et al*., 2000; Yeung and Ruzzo, 2001; Landgrebe *et al*., 2002; Méndez *et al*., 2002).

Fisher's linear discriminant analysis (FLD) is a linear dimensionality reduction technique that is optimal in terms of maximizing the separation amongst these classes. Where PCA seeks directions that are efficient for representation, FLD seeks directions that are efficient for discrimination. Hence, FLD determines a set of projection vectors which simultaneously maximize the scatter between classes and minimize the scatter within each class, and which maximize the separability of the data (Duda *et al*., 2001). The projection vector of FLD can be obtained by solving the following optimization problem:

$$\max_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T\mathbf{S}_\mathbf{B}\mathbf{w}}{\mathbf{w}^T\mathbf{S}_\mathbf{W}\mathbf{w}} \tag{6}$$

where $\mathbf{S}_\mathbf{B}$ is the between-class scatter matrix, $\mathbf{S}_\mathbf{B} = \sum_{i=1}^{c} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$, $\bar{\mathbf{x}}_i$ is the mean vector for class i, and $\bar{\mathbf{x}}$ is the total mean vector (Duda *et al*., 2001). With FLD vectors determined, each sample can then be classified in this reduced FLD space using discrimination analysis. Several researches are used to create a linear projection of gene expression measurements that maximizes the separation of different sample classes and to develop the classification method which used a distance measure within a FLD space (Cho *et al*., 2002; Hwang *et al*., 2002; Stephanopoulos *et al*., 2002).

1.2 Cancer classification by machine learning

The purpose of supervised gene expression data analysis is to construct well-performing classifiers using machine learning algorithms such as linear discriminants, decision trees or support vector machines, which assign predefined classes to a given expression profile (Alizadeh *et al*., 2000). In this paper, two classification methods of linear discriminant function and support vector machine are used to classify the type of a cancer.

A linear discrimination function (LDF) that is a linear combination of the components of $\mathbf{x}$ can be written as

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + \mathbf{b} = 0 \tag{7}$$

where $\mathbf{w}$ is the weight vector and $\mathbf{b}$ is the bias. Given the sets of the input vector to train the classifier, the training process involves the adjustment of the weight vector $\mathbf{w}$ in such a way that the two classes (w_1 and w_2) are linearly separable (Haykin, 1999; Duda *et al*., 2001).

On the other hand, support vector machines (SVM), which are a kind of supervised machine learning techniques, have been shown to perform well in multiple areas of biological analysis, including the evaluation of microarray gene expression data. While the linear discriminant analysis described above can produce linear decision boundaries for the classification, SVM produces nonlinear boundaries as a result of generating linear decision boundaries in the feature space (Vapnik, 1995; Hastie *et al*., 2001). SVM has demonstrated the ability to not only correctly separate entities into appropriate classes, but also to identify instances whose established classification is not supported by the data (Brazma and Vilo, 2000; Brown *et al*., 2000; Furey *et al*., 2000; Shipp *et al*., 2002).

2. Supervised Clustering and Classification by Machine Learning Algorithms

In this paper, a new supervised hierarchical clustering algorithm is proposed including a new metric that uses additional information available from the gene selection with the DPLS method. VIP values, representing the importance of the genes in the DPLS, can be considered as a good information source for increasing the clustering efficiency and interpretation capability of a clustering algorithm. It is reasonable to assume that the weights of the genes in the clustering

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 genes with respect to the class label.

A new weighted Euclidean distance ($d_{ij}^{(w)}$) is proposed as a distance metric using the normalized VIP weights as follows:

$$d_{ij}^{(w)} = \sqrt{\left(\sum_{k=1,...N} w_k (x_{ik} - x_{jk})^2 \right)} \quad (8)$$

where $\mathbf{w} = (w_1, w_2, ..., w_k)$ is the feature-weight vector with the normalized VIP values and N is the number of samples. Therefore, the normalized weights which are based on the VIP weights are assigned to each gene for indicating the importance of those genes (Questier *et al.*, 2005). Their weights are the importance degree corresponding to each feature. The larger w_k (VIP_k) is, the more important the k-th gene is in the hierarchical clustering. When $\mathbf{w}$ is (1, ..., 1), the space $\{\|d_{ij}^{(w)}\| \le r\}$ is a hypersphere with radius r in the well-known Euclidean space (called the original space). In the original space, $d_{ij}^{(w)}$ is then denoted by d_{ij} and the supervised hierarchical clustering would reduce to an unsupervised hierarchical clustering algorithm. When $\mathbf{w} \neq$ (1, ..., 1), it means that the axes of the hypersphere would be extended or shrunk in accordance with the value of w_k. Thus in this case the space $\{\|d_{ij}^{(w)}\| \le r\}$ is a hyper-ellipse, and the lower the value of w_k is, the higher the flattening extent of the ellipse is for that dimension. It is well known that an appropriate assignment of feature-weights can improve the performance of feature-weighted (supervised) hierarchical clustering algorithms (Wang *et al.*, 2004). This weighted distance can be easily proven to satisfy four requirements of a generic metric: nonnegativity, reflectivity, symmetry and triangle inequality.

According to Johnson and Wichern (1992), the following are the steps in the suggested agglomerative supervised hierarchical clustering algorithm for a group of N samples.

1. Start with N clusters, each contains a single entity and an $N \times N$ symmetric matrix of the weighted distance $\mathbf{D} = \{d_{ij}^{(w)}\}$.
2. Search the weighted distance matrix for the nearest (most similar) pair of clusters. Let the distance between the "most similar" clusters U and V be d_{UV}.
3. Merge cluster U and V. Label the newly formed cluster (U V). Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (U V) and the remaining clusters.
4. Repeat steps 2 and 3 a total of (N – 1) times. Record the identity of clusters that are merged and the levels at which the mergers take place.

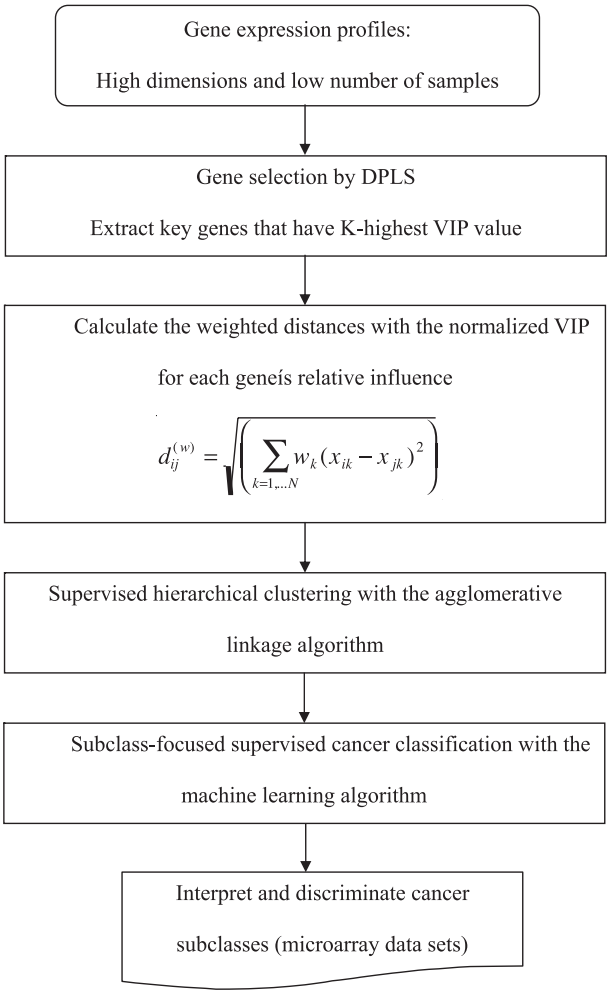


Fig. 1 Schematic diagram of the gene selection, the supervised clustering, and the subsequent subclass-focused cancer classification using gene expression profiling

Figure 1 shows a schematic diagram of the supervised gene selection, clustering and classification by supervised machine learning algorithms for discriminating cancer subclasses. First, the key genes are selected using the VIP information of the DPLS model. Secondly, a weighted Euclidean distance is calculated for the proposed supervised clustering using the VIP weights. Thirdly, multivariate analysis and supervised clustering for cancer subclasses are used to interpret gene expression patterns, to classify the sample into a subclass, or to predict the clinical results of treatment. Because the proposed method makes use of the DPLS method in both supervised gene selection and supervised clustering, we expect to obtain synergistic effects of the supervised knowledge that is responsible for selecting the key genes, by sequentially using the influence of the key genes into a supervised hierarchical clustering and classification method. Supervised machine learning algorithms are applied to three

Table 1 Classification of the 72 microarrays according to leukemia subtype and, when available, information on the outcome of the leukemia treatment (Golub *et al*., 1999)

No.	ALL/AML	Subclass	Treatment result	No.	ALL/AML	Subclass	Treatment result
1	ALL	B-cell		38	AML	M1	Success
2	ALL	T-cell		39	ALL	B-cell	
3	ALL	T-cell		40	ALL	B-cell	
4	ALL	B-cell		41	ALL	B-cell	
5	ALL	B-cell		42	ALL	B-cell	
6	ALL	T-cell		43	ALL	B-cell	
7	ALL	B-cell		44	ALL	B-cell	
8	ALL	B-cell		45	ALL	B-cell	
9	ALL	T-cell		46	ALL	B-cell	
10	ALL	T-cell		47	ALL	B-cell	
11	ALL	T-cell		48	ALL	B-cell	
12	ALL	B-cell		49	ALL	B-cell	
13	ALL	B-cell		50	AML	M4	Failure
14	ALL	T-cell		51	AML	M2	Failure
15	ALL	B-cell		52	AML	M4	Success
16	ALL	B-cell		53	AML	M2	Success
17	ALL	B-cell		54	AML	M4	
18	ALL	B-cell		55	ALL	B-cell	
19	ALL	B-cell		56	ALL	B-cell	
20	ALL	B-cell		57	AML	M2	
21	ALL	B-cell		58	AML	M2	
22	ALL	B-cell		59	ALL	B-cell	
23	ALL	T-cell		60	AML	M2	
24	ALL	B-cell		61	AML	M1	
25	ALL	B-cell		62	AML		
26	ALL	B-cell		63	AML		
27	ALL	B-cell		64	AML		
28	AML	M2	Failure	65	AML		
29	AML	M2	Failure	66	AML		
30	AML	M5	Failure	67	ALL	T-cell	
31	AML	M4	Failure	68	ALL	B-cell	
32	AML	M1	Failure	69	ALL	B-cell	
33	AML	M2	Failure	70	ALL	B-cell	
34	AML	M2	Success	71	ALL	B-cell	
35	AML	M1	Success	72	ALL	B-cell	
36	AML	M5	Success				
37	AML	M2	Success				

microarray data sets obtained from leukemia, breast, and colon cancer patients to establish a relationship between the microarray data and expression-based cancer subclasses as well as patient treatment outcome.

3. Results and Discussion

3.1 Leukemia gene expression profiles

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 normally generates cells called blasts that develop (mature) into several different types of blood cells with specific tasks in the human body. Acute leukemia data set can be classified into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). 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). 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.

The leukemia data set and all details with respect to the methods used to collect the data are described in the paper of Golub *et al*. (1999). The data set consists of a set of high-density oligonucleotide microarrays (Affymetrix) with probes of 7129 human genes, and was obtained from 72 patients. 47 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 (Golub *et al*., 1999). **Table 1** provides some information on the leukemia subclasses to which each of the leukemia microarray data sets belongs. Moreover, the result of the subsequent medical treatment (success or failure) is provided for a limited number of samples. The gene expression profiles of the original data set are represented as log10 normalized expression values, such that overall intensities for each chip are equivalent. To remove systematic sources of variation in the microarray experiments, the expression level of each gene was normalized to have a zero mean and a standard deviation of one (Yang *et al*., 2002).

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

3.2 Supervised clustering of leukemia data set

3.2.1 Supervised clustering between ALL and AML

The DPLS method was used for selection of the genes that are most suited to discriminate between AML and ALL in the training data set of Golub *et al*. (1999), where the response variable **Y** was either 0 (AML) or 1 (ALL). Out of the 7129 available genes in the expression data, the 23 genes which are most correlated with the leukemia classification into ALL or AML were selected on the basis of the VIP value resulting from applying DPLS. The cross validation method was used to determine the number of relevant genes to be retained. This method reconstructs a DPLS model with a minimal number of genes until the classification performance of DPLS in a cross validation does not decrease. This procedure finally resulted in the selection of 23 genes as the minimum number of genes that provides a good classification performance. In contrast to Golub *et al*. (1999), who selected 50 relevant genes, the approach used here results in the selection of a significantly lower number of relevant genes. The DPLS-based gene selection method assigns high rankings to zyxin, leukotriene (C4 synthase gene), leptin, CD33 antigen, FAH, 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 (Cho *et al*., 2002). In addition, CD33 is located in chromosome 19q13.3, and has been developed for targeted antibody therapy to kill leukemia AML cells (Golub *et al*., 1999; Thomas *et al*., 2001; Bicciato *et al*., 2002; Cho *et al*., 2002).

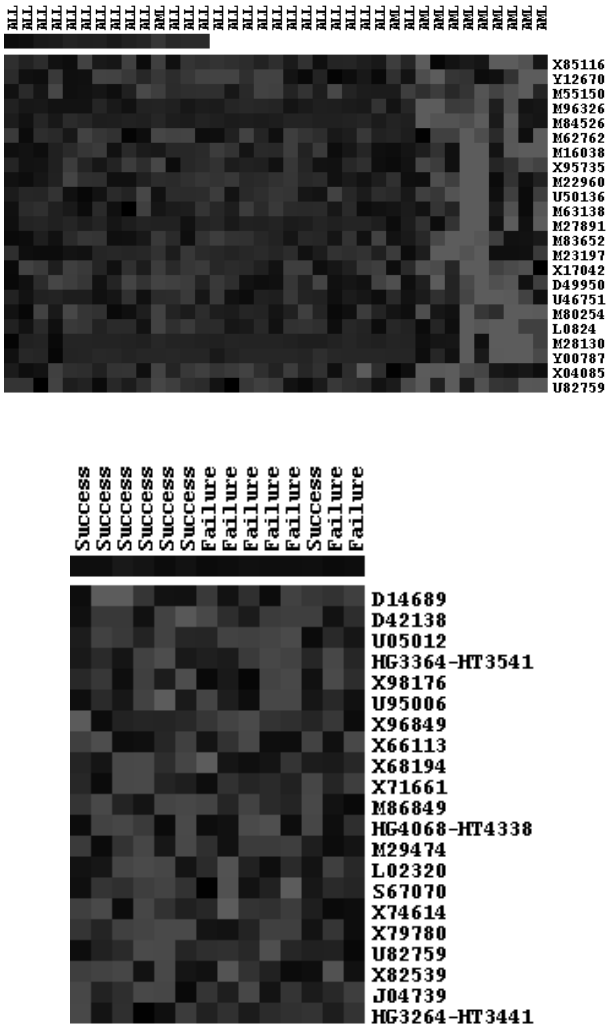
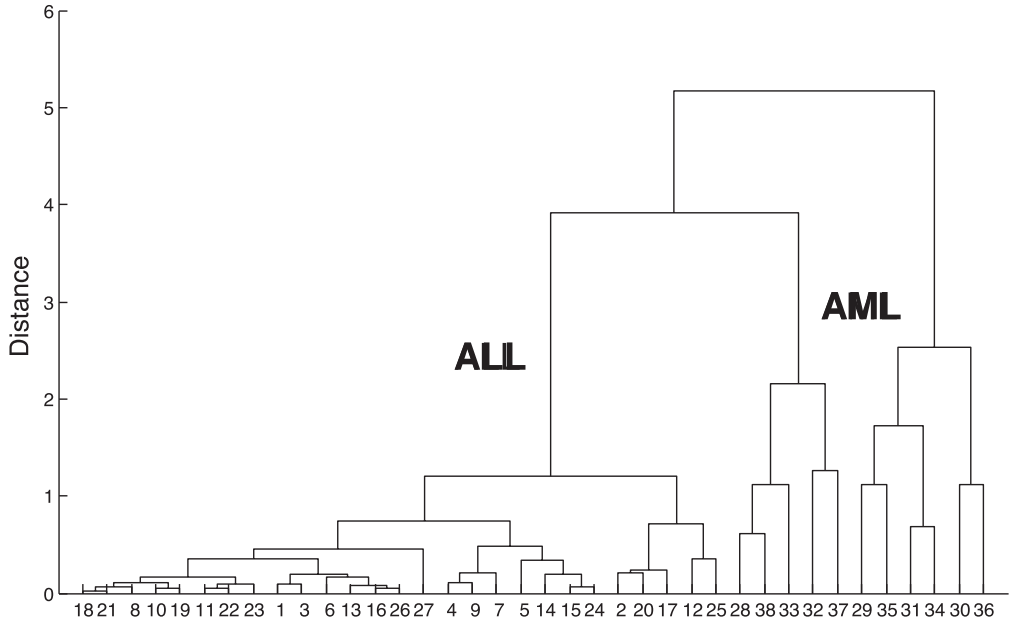


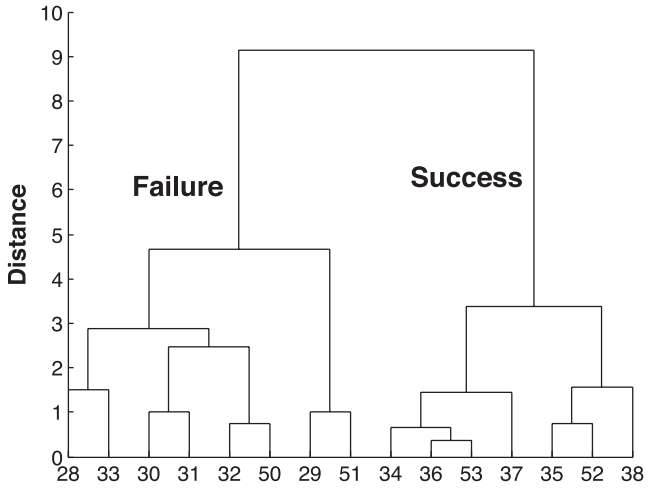
Fig. 2 Gene expression heat maps of a leukemia data set based on (a) the 23 selected genes most relevant for discrimination between ALL and AML and (b) the 21 selected genes most relevant for discrimination between success and failure of AML leukemia treatment

Figure 2(**a**) shows gene expression maps of a leukemia data set based on the 23 selected genes most relevant for discrimination between ALL and AML, where we used CLUSTER and TREEVIEW software, which are both publicly available at http://rana.lbl.gov. The figure confirms that the expression of the selected genes is significantly different for ALL and AML samples, and that expression of the individual genes is rather similar within each class. From this figure, we can conclude that the genes selected via VIP are discriminatory.

Supervised hierarchical clustering with the agglomerative linkage algorithm was applied to the 38



(a)



(b)

Fig. 3 Dendrogram of supervised clustering for the leukemia training data set based on (a) the 23 genes most relevant for discrimination between ALL and AML and (b) the 21 genes most relevant for discrimination between success and failure of AML leukemia treatment

samples of the training data set, in order to cluster the samples on the basis of their weighted similarities over the selected 23 genes. In the dendrogram (**Figure 3**(**a**)), two clusters appear, corresponding to ALL and AML respectively. There is no discrimination error for the supervised clustering. On the other hand, when no information on the relative importance of each gene for leukemia class distinction was used in the clustering, the unsupervised clustering analysis shows a single misclustered sample (sample 35), which is AML but was clustered as ALL. This confirms that the supervised (weighted) hierarchical clustering analysis can improve the clustering performance for discriminating ALL and AML subclasses, since the normalized weights of the VIP can give relative contribution values to each gene for discrimination of subclasses.

3.2.2 Supervised clustering for ALL subclass (B-cell and T-cell) ALL can be further classified into 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 to 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). The DPLS

method combined with the cross validation method resulted in the selection of 15 genes that allow discriminating between T-cell ALL (T-ALL) and B-cell ALL (B-ALL). The DPLS response variable **Y** is 0 (T-ALL) or 1 (B-ALL).

The 15 genes were examined for chromosomal localization using NCBI LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink). Almost all genes are mapped to regions that have been previously associated with ALL chromosomal abnormalities, including the T-cell antigen receptor (X03934, 9p56), TCRB (X00437, 7q34) (CD47, X69398 3q13), CD7 (D00749, 7q34) and TCF7(X59871, 5q31). The results thus suggest that the 15 genes selected via DPLS as being most relevant for discriminating between B-ALL and T-ALL subclasses are biologically relevant as well. The subsequent supervised hierarchical clustering for the training data set using the expression information of the 15 selected genes resulted in correct clustering for all samples, confirming that the selected genes can be used to discriminate B-ALL and T-ALL subclasses.

3.2.3 Supervised clustering between AML subclasses

The original French-American-British (FAB) system for detemining leukemia subtype was only based on the appearance of leukemic cells under the microscope after routine processing or cytochemical staining. AML can be classified into six subtypes, designated M1, M2, M3, M4, M5, and M6. Although patients tend to be classified into either the M2 or M4 subtype under the FAB system, it is difficult for most doctors to discriminate sharply between these subtypes. 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. Clearly, correct identification of the AML subtype is very important to the subsequent clinical treatment step (Golub *et al*.,1999).

Among the 25 AML patients, data of 20 patients were used as a training data set, (4 M1 cases: samples 32, 35, 38, 61; 10 M2 cases: samples 28, 29, 33, 34, 37, 51, 53, 57, 58, 60; 4 M4 cases: samples 31, 50, 52, 54; 2 M5 cases: samples 30, 36). The DPLS method was again used for determining the minimum number of relevant genes that allow discriminating between the AML subclasses. The response matrices (**Y**) were [1 0 0 0] for M1, [0 1 0 0] for M2, [0 0 1 0] for M4 and [0 0 0 1] for M5 respectively. Many of the genes in the top 25 genes relevant for AML subclass discrimination (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 (*TFIIEβ*), 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 *et al*., 2001).

Unsupervised and supervised clustering were applied to the training data set for the 25 selected genes most relevant for discrimination between AML subclasses (M1, M2, M4 and M5). For both clustering methods, there is one misclustered sample: Sample 51 is M2 but was clustered as M1.

3.2.4 Supervised clustering for clinical outcome of AML patients (failure and success) Relating gene expression patterns to the clinical outcome of cancer treatment is a key issue in cancer genetics. 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. Gene selection for the prediction of the clinical output of AML treatment (success or failure) was performed using DPLS. Among the 25 AML patients of Golub *et al*. (1999) with know clinical outcome of leukemia treatment, 15 samples formed a training data set (7 patients survived: 34–38 and 52–53; 8 patients died during treatment: 28–33, 50 and 51 (see Table 1). The response variable **Y** was 0 (success) or 1 (failure).

A subset of 21 genes was selected for discriminating between failure and success of clinical treatment of AML patients, as a result of applying DPLS with the crossvalidation method. The chromosomal locations of the 21 identified genes were checked in the NCBI, because chromosomal abnormalities are prevalent in leukemia patients and often have prognostic implications (Thomas *et al*., 2001). Almost all genes among the selected 21 genes have been identified previously as containing abnormalities in AML or another form of leukemia. Most of the genes reported by Lyons-Weiler *et al*. (2003) are also found in our DPLS gene set (*HoxA9*, *PIG-B*, *MACH-alpha-2* protein, *BPI* Bactericidal/permeability increasing protein, Autoantigen *PM-SCL*, *ERGIC-53* Protein, and so on). **Figure 2**(**b**) shows a heat map of the leukemia gene expression data based on the 21 selected genes most relevant for discrimination between success and failure of AML leukemia treatment. It illustrates that the selected genes have a significant expression difference for AML leukemia treatment success and failure. The figure thus indicates that the selected genes can be used as a set of diagnostic genes for discrimination between the treatment results of AML patients.

Figure 3(**b**) shows the dendrogram resulting from applying the supervised hierarchical clustering analysis to the training data set using the selected 21 genes. This clustering analysis divides the treatment failure and success in two distinct groups, which perfectly match with the recorded treatment results. The location of all successful treatments and all treatment failures in two separate groups indicates that the candidate

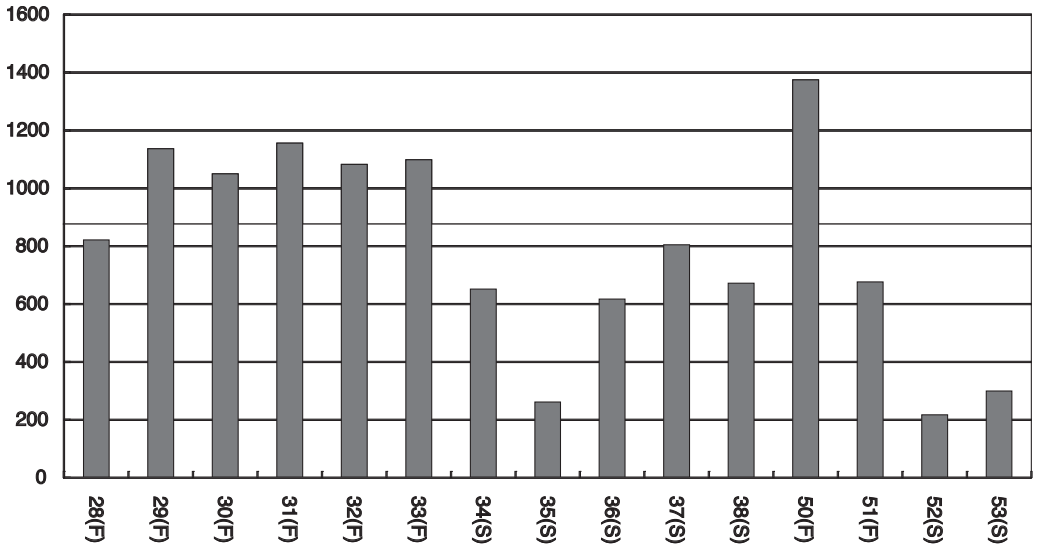


Fig. 4 Comparison of gene expression profiles of *HoxA9* in the 15 AML patients with known clinical outcomes (success = S; failure = F)

genes selected via VIP can be used successfully as a set of diagnostic genes for discrimination between the treatment results of AML patients. Similarly to previous cases, the clustering appears to be improved by inclusion of VIP weights.

The gene selection indicated that *HoxA9* was the most relevant gene for discriminating between successful and failing leukemia treatment. Overexpression of *HoxA9* would presumably result in an overproduction of leukocytes and lymphocytes. Indeed, the injection of retrovirally engineered primary bone marrow cells that overexpressed 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 ($P < 0.1$). **Figure 4** shows the gene expression profiles of *HoxA9* in the 15 AML patients with known clinical outcomes (success = S; failure = F), where S means the patient survived after the treatment and F means the patient died after the treatment.

Among these patients, those with poor treatment outcomes showed increased expression of *HoxA9*. Biotechnological advances, such as gene expression profiling via DNA microarrays, allow researchers to enlarge their understanding of the mechanisms underlying diseases. The DNA microarray technology is useful when applied to RNA extracted from tissue samples: The resulting data allow discriminating between various subtypes of leukemia, which is necessary for the accurate diagnosis and treatment of patients. From the results demonstrated in this paper, we can conclude that the gene selection method via VIP can be used to select key genes for discriminating leukemia types and subtypes. The method also allowed successful prediction of medical outcome of leukemia treatment using gene expression data. Moreover the supervised clustering can improve the clustering performance for discriminating leukemia subclasses, compared to unsupervised clustering.

3.3 Supervised Classifications by Machine Learning

With the information on the most relevant features in a gene expression data set available, a following step is to build a robust cancer classifier capable of correctly predicting the sample labels from the available expression profiles. Supervised machine learning techniques are well-suited for this purpose. Two feature selection methods are compared: PCA as an unsupervised feature selection method and FLD as a supervised feature selection method. Two supervised classification methods, LDF and nonlinear SVM are subsequently applied to classify leukemia gene expression samples. This paper focuses on two case studies: Classification of samples into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) on the one hand, and classification of clinical treatment outcomes for AML patients (treatment success or failure) on the other hand. These two case studies are selected since they are considered to be more important for the classification of leukemia patients than the other case studies considered in the first part of this paper.

3.3.1 Classification of ALL and AML samples To interpret and correctly classify gene expression data obtained from ALL and AML patients, PCA was first

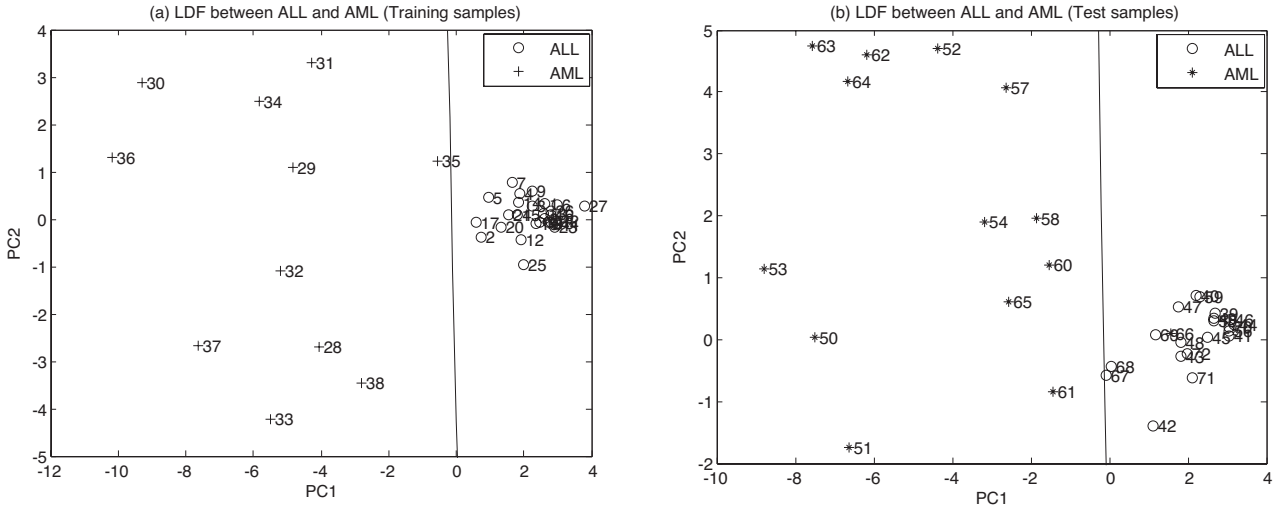


Fig. 5 Classification result of AML and ALL classes using linear discriminant functions: (a) training samples (38 patients) and (b) test samples (34 patients)

applied to obtain a further dimension reduction of the 23 most relevant genes retained by Yoo *et al*. (2005) in order to interpret and avoid overfitting problems, since the gene expression data are too related to the response value. LDF was subsequently applied as a classifier.

The dimension reduction obtained with PCA allowed to visualize the patterns of ALL and AML in the leukemia data set, and to improve the classification performance. Indeed, two PCs which capture about 74.3% of the variation in the 23 genes were found to be adequate based on the cross-validation of the prediction residual sum of squares (PRESS). The LDF classification results for AML and ALL samples are provided in **Figure 5**. Figure 5(a) illustrates the classification result of the LDF for the training data set (38 samples), where ALL and AML samples are well separated. Figure 5(b) shows the classification results of 34 test patients in the two-dimensional space formed by the first two PCs. The test data set was classified with 97% accuracy (33/34 patients) using LDF. The only misclassified sample corresponds to patient 66, which was classified as ALL but actually labeled AML. Reclassifying the samples using the original 23 genes selected in the first paper, also showed one misclassification result for sample 66 using LDF. Several investigations indicated that the leukemia data set of Golub *et al*. (1999) contains at least one sample including patient 66 that is mislabeled and patient 66 has unusually low gene expression levels compared to other AML patients. Yoo *et al*. (2005) showed that the contributions of most genes in the 66th patient are negative, contrary to the behavior of the other AML patients. In particular, the top-ranked gene, Zyxin, shows an abnormally low expression level for patient 66. It means that there is actually no classification error for the 34 test data since sample number #66 is known as the mislabeled sample in the leukemia data set. Thus, it can be concluded that a further dimension reduction by PCA enables extraction of meaningful features that permit distinguishing between ALL and AML. This classification result is superior to that of Golub *et al*. (1999), who obtained a total of five misclassifications from applying a weighted voting scheme, a variation of a diagonal linear classifier. The result is also better than that of Liang and Kachalo (2002), who achieved three classification errors by a linear classifier.

In developing a nonlinear classifier like SVM, the most important thing is the extraction of appropriate features, that is which features are retained from the original inputs, in order to avoid the overfitting problem (Schölkopf *et al*., 2000). In order to see the effect of the feature extraction on the classifier performance, four different classifiers were used on the data: Linear and nonlinear SVM were applied to the original gene data (23 genes), selected by Yoo *et al*. (2005) and to two reduced scores resulting from applying FLD to the original data. Each classifier uses the SVM algorithm to define a hyperplane that best separates the training samples into two classes, ALL and AML. In this paper, the radial basis function $\exp(-\|x - y\|^2/c)$ is used as a SVM kernel function to capture the nonlinearity. The width of a Gaussian function $c = 1.5\mu$ is selected, as suggested by Cremers *et al*. (2003), in order to get a smooth energy landscape, where μ is the average distance between two neighboring data points. Data normalization is used for kernels to improve the condition number of the Hessian in the optimization routine. The parameter C which places an upper bound on the Lagrange multipliers is set to 1 for implementing linear and nonlinear SVM, and for reducing the number of support vectors.

Table 2 Comparison of classification results of several classifiers obtained by applying FLD and SVM for discriminating between ALL and AML (*Note: sample number #66 is known as a mislabeled sample in the leukemia data set)

Method	Classifiers	No. of SV	Training error	Validation error
Method 1	Original data + linear SVM	5	0	2 (#61, 66)
Method 2	Original data + nonlinear SVM	35	1(#35)	1 (#66)
Method 3	FLD + linear SVM	2	0	1 (#66)
Method 4	FLD + nonlinear SVM	2	0	1 (#66)

Table 2 shows the classification performance of the four classifiers for identifying ALL and AML labels. Linear and nonlinear SVM have very similar classification results for distinguishing between ALL and AML, both for the original features and the FLD scores. The training samples are perfectly classified, except for method 2, the nonlinear SVM applied to the original features. The classification performance evaluation for the test samples shows 1 to 2 errors for all four classifiers on a total of 34 test samples. Compared to the other methods, the result of method 1 (linear SVM applied to the original features) was not good, with 2 misclassified test samples. It indicates that the original data (23 genes) may contain non-separable signals or be corrupted by a high noise. Note also that the number of support vectors for method 2 is 35 while only 38 training samples were used. This means that the original 23 genes may contain a lot of noise and can easily be overfitted. A high number of support vectors for a data set forms indeed an indication that the SVM model may be overfitted, and that significant misclassification rates can be expected for the training and test samples. The classification results of methods 3 and 4, which are the linear and nonlinear SVM applied to two FLD score vectors, are quite similar. Sample 66 is misclassified with 2 support vectors. The features extracted via FLD make it a simple model, in which the number of support vectors for the classifier decreases to only two. In spite of this small number of support vectors, the classification results for both the training and test samples are good. Note that sample number #66 is known as the mislabeled sample in the leukemia data set. It means there is no classification error of SVM for the test data except the mislabeled sample #66. This demonstrates the fact that a dimension reduction (i.e. feature extraction) by FLD can improve the generalization performance of a classifier such as SVM.

To further evaluate the performance of the proposed method, we compared our method with previously developed methods which were applied to the same leukemia microarray data set. In general, it is somewhat difficult to directly compare these methods because they each use a different criterion. We compared their classification performance using the number of the misclassification samples. **Table 3** compares the misclassification results of 34 test set using some previously published results, such as decision tree learner J48 (Weka's implementation of C4.5), simple Bayesian classifier or naive Bayes, sequential minimal optimization(SMO)-wrapper, emerging patterns (Li and Wong, 2002), SVM (Fuery *et al*., 2000), voting machine (Golub *et al*., 1999), maximal margin linear programming (MAMA, Antonov *et al*., 2004), and projective adaptive resonance theory (PART, Takahashi *et al*., 2005). As one of the most well-known results, the previous result of the voting machine with 50 genes by Golub *et al*. (1999) can correctly predict 29 samples in the test set with 4 misclassified samples. The proposed method can correctly predict the test samples with a relatively low number of selected genes (23) except for a single misclassified sample (#66), which is known as a mislabeled sample and may influence the error rate (Chow *et al*., 2001). The comparison of the method proposed in this paper with previously published results thus demonstrates that classification performance of the proposed method is equivalent or better than results reported in the literature.

Table 3 The comparison of classification results for ALL/AML classification in a test set of 34 samples

Method	No. of genes	Misclassification results
J48[a]	1	3
Naive Bayes[a]	1	3
SMO-CFS[a]	1	3
SMO-wrapper[a]	2	4
Emerging patterns[b]	1	3
SVM[c]	25–1000	2–4
Voting machine[d]	50	5
MAMA[a]	132–549	0

[a]Wang *et al*. (2005); [b]Li and Wong (2002); [c]Furey *et al*. (2000); [d]Golub *et al*. (1999); [e]Antonov *et al*. (2004); [f]Takahashi *et al*. (2005)

3.3.2 Classification of clinical outcome of AML patient treatment (success or failure) The FAB system for classifying AML subtypes was originally only based on the morphological states of cells under the microscope, and has later on been extended with criteria based on immune markers and cytogenetic abnormalities. Correct determination of the AML subtype is important, since different subtypes will respond

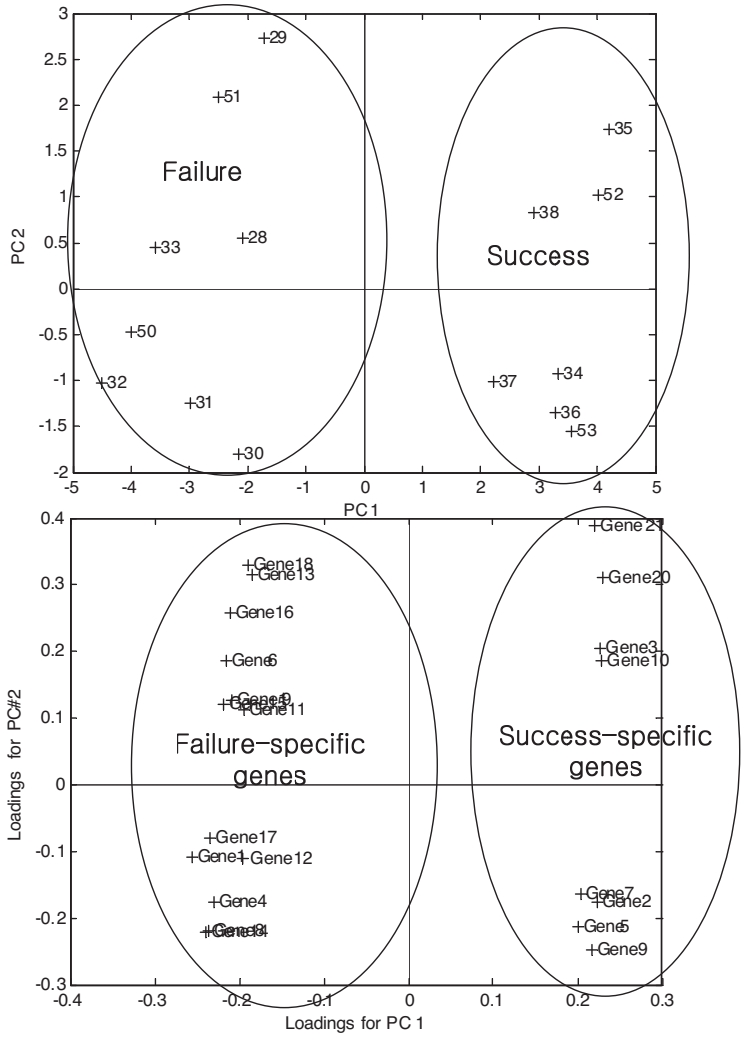


Fig. 6 Interpretation and classification result using PCA for 15 AML patients (8 failures and 7 successes): (a) score plot and (b) loading plot

differently to medical treatment. Gene expression data also contain information that can elucidate the success or failure of leukemia treatment. Designing a suitable classifier thus allows predicting the clinical outcome of leukemia patient treatment. The 21 most relevant genes for discriminating between failure and success of clinical treatment of AML patients selected in the first part of this paper were used as a starting point for the development of a classifier. PCA was used to reduce the data dimensionality, and to interpret the clinical outcome of AML patient treatment. Two PCs were used, which captured about 63.6% of the variation in the 21 genes.

The score and loading plots of the 15 AML patients included in the training set (8 failures and 7 successes) were examined for determining the correlation between the selected genes and the clinical outcome of AML patient treatment (**Figure 6**). The plots clearly demonstrate that the selected 21 genes can exactly discriminate the clinical outcome of AML patients, and that PCA can extract the key feature components. The loadings plot in Figure 6(b) can be used to establish how the 21 genes are interrelated. The shape of the loading plot is closely connected with the pattern of the score plot in Figure 6(a), and shows how the 21 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 hand side and genes that correlate with treatment failure appear on the left hand side. Almost all of the genes in each gene group have common expression patterns, that is, group-specific regulation patterns known as co-regulation patterns. It means that the expression of each group is highly elevated only in one sample class, and down-regulated in the other classes (Stephanopoulos *et al*., 2002). This result is notable in that these genes may be considered marker genes related to the clinical outcome of AML patient treatment.

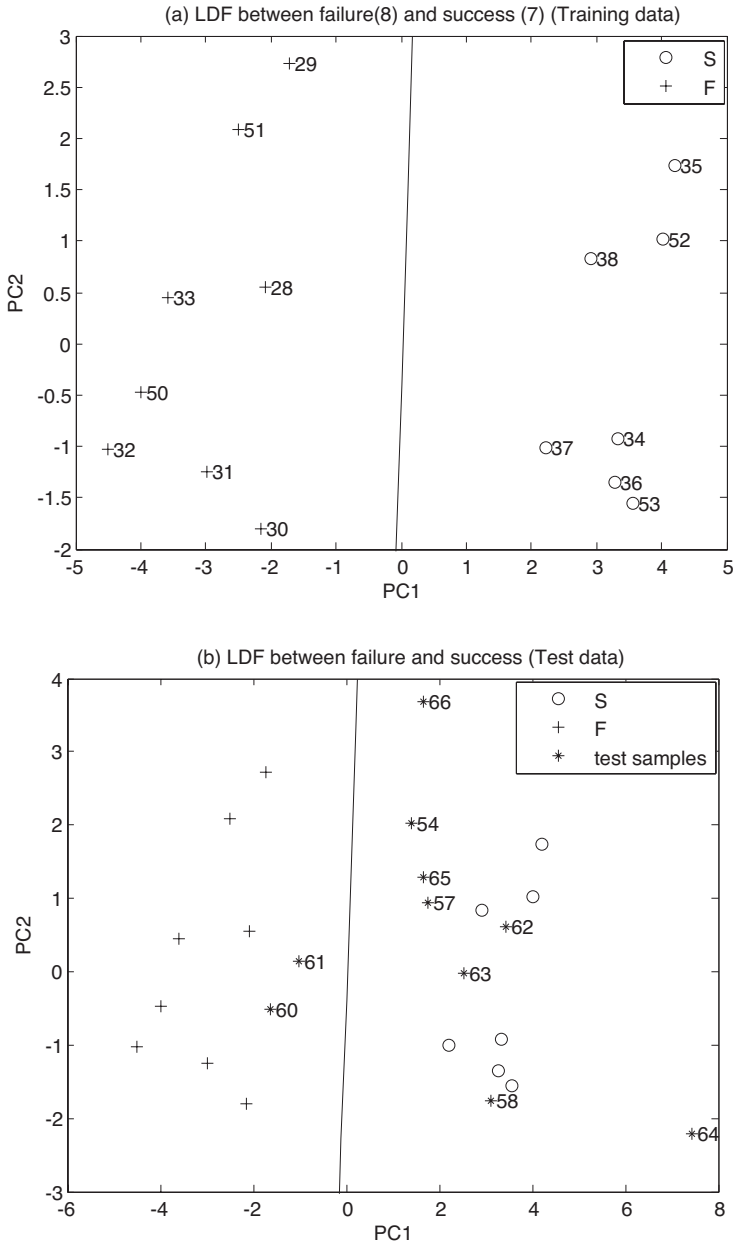


Fig. 7 Classification results of LDF for AML patients: (a) 15 training samples (8 failures and 7 successes) and (b) 10 test samples of unknown leukemia treatment outcome

Figure 7 shows the classification results obtained by combining PCA and LDF for all 25 AML patients, including the 15 training samples (8 treatment failures, 7 successes), and 10 test samples (patient 54, 57, 58, 60–66). Samples are plotted in the space spanned by the first two PCs. Figure 7(a), after applying the linear classifier in the two PC space, illustrates that the information contained in the 21 genes provides excellent separation for the 15 training sample AML patients with respect to the clinical outcome of the treatment. Figure 7(b) depicts the classification results of PCA and LDF for the 10 AML patients in the test data set (54, 57, 58, 60–66), whose clinical outcome was not specified by Golub *et al*. (1999). The classification results point towards successful treatment for eight AML patients (#54, 57, 58, 62, 63, 64, 65, and 66). Two AML patients (#61 and 62) are predicted not to survive treatment. Thus, the classifier is able to predict the clinical outcome of AML patients, but the performance of the classifier on the test samples cannot be evaluated since the appropriate information on patient survival is not available.

Table 4 represents the prediction results of several SVM classifiers for the 10 AML patients of the test data set, whose clinical outcome was not specified by Golub *et al*. (1999). All classifiers, linear as well as

Table 4 Prediction results of 10 AML test patients using several classifiers obtained by applying FLD and SVM, where S indicates the survival of the patient after treatment and F indicates the death of the patient after treatment)

Sample No.	Original data + Linear SVM	Original data + nonlinear SVM	FLD + Linear SVM	FLD + nonlinear SVM
54	S	S	S	S
57	S	S	S	S
58	S	S	S	S
60	F	F	F	F
61	F	F	F	F
62	S	S	S	S
63	S	S	S	S
64	S	S	S	S
65	S	S	S	S
66	S	S	S	S

Table 5 The comparison of classification results for breast and colon cancer data sets

Data set	Method	Misclassification results
Breast cancer data set	Bayesian variable selection[a]	0
	Kernel FDA[b]	5
	Kernel Fisher FDA[d]	0
	Method proposed in this paper	0
Colon cancer data set	Bayesian classification method[c]	2.90
	Kernel Fisher FDA[b]	2.57
	Kernel Fisher FDA[d]	2.15
	Method proposed in this paper	2.11

[a]Lee *et al*. (2003); [b]Cho *et al*. (2003); [c]Li *et al*. (2002); [d]Cho *et al*. (2004)

nonlinear classifiers, both applied on the original gene expression data (21 genes) as well as on two FLD score vectors, showed exactly the same clinical outcome prediction of AML patients. Eight AML patients (samples 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 conclusion is that a dimension reduction by FLD enables extracting the meaningful features to discriminate between success and failure of AML treatment. Based on the present findings with regard to the link between expression of certain genes and clinical outcome of AML treatment, determining the specific genes combined with the proposed classifiers would allow predicting relapse in leukemia patients. Although 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.

3.4 Supervised classification of breast and colon cancer data sets

To evaluate the performance of the proposed supervised classification method, two additional data sets are selected—a breast cancer microarray data set and a colon cancer microarray data set—and the proposed method is compared with three previously developed methods. The breast cancer data set is used for the discrimination between *BRCA1* mutation, *BRCA2* mutation and other mutations (data set consisting of seven *BRCA1* mutation samples and eight *BRCA2* mutation samples and sporadic samples) with 3226 genes and 22 samples (Hedenfalk *et al*., 2001). The colon cancer data set is used for diagnosis of cancer patients and consists of 2000 probes and 62 samples (40 cancer tissues and 22 normal tissues, Alon *et al*., 1999). Although it is somewhat difficult to compare these methods because they each use a different criterion, the number of misclassifications is used for the comparison. **Table 5** shows the comparison results with Bayesian variable selection (Lee *et al*., 2003), kernel Fisher discriminant analysis (FDA, Li *et al*., 2002; Cho *et al*., 2003, 2004). The gene set that was selected by Cho *et al*. (2004) is used in this paper.

For the breast cancer data set, the proposed method shows a satisfactory classification result in Table 5. While Cho *et al*. (2003) produced 3 misclassification samples over three models, the results of our method and results reported by Lee *et al*. (2003) and Cho *et al*. (2004) show zero misclassification results. Note that

the approach of Lee *et al*. (2003) is based on a quite complex method which is composed of Bayesian mixtures and markov chain monte carlo computation. From this result, we can conclude that the proposed method performs well and is much simpler and thus easier to use than the methods of Lee *et al*. (2003) and Cho *et al*. (2003, 2004).

For the colon cancer data set, Li *et al*. (2002) previously analyzed it using the average performance over 100 random partitions into 50 training and 12 test samples. As shown in Table 5, the proposed method—reaching classification using a limited set of informative genes which are specific to a certain type of cancer—shows a minimum average test error that is lower than the error reported by Li *et al*. (2002). The selected genes in the colon cancer data set used for classification with the proposed method contain the vascular endothelial growth factor (*VEGF*, IMAGE ID:47326). The clinical studies show that VEGF is a dominant angiogenic factor in human colorectal cancer and is associated with the formation of metastases and poor prognosis (Cho *et al*., 2004).

These results lead to the following conclusion. First, the classification result which can exactly classify the tumor type is mainly dependent on the selected genes. Research by Dettling and Buhlmann (2002) also shows that a supervised clustering algorithm can identify functional groups of interacting genes that have high explanatory power for the given tumor type, which in turn can be used to accurately predict the class labels of new samples. Second, as noted by Kulkarni *et al*. (2005), supervised clustering with DPLS information of the tumor types of the tissues makes local transformations with supervised translations in the gene expression data. It changes the representation of the data whose class overlap is decreased slightly as compared to their original data distribution.

Conclusions

In this paper, we develop a supervised framework for the gene selection, clustering, and classification of microarray gene expression profiles, thus allowing discrimination between cancer subclasses. First, the marker genes which have great classification ability for given cancer types are selected. Second, supervised clustering using the valuable weights information of DPLS was suggested to subsequently group the tumor samples into different classes, where the normalized weights of VIP can give relative contribution values for the discrimination of subclasses. Third, supervised linear and nonlinear classification methods were applied to three microarray data sets of (leukemia, breast, and colon cancer) to predict and classify the tumor samples according to their membership to particular tumor classes. Supervised machine learning algorithms enable the classification of leukemia subtypes solely on the basis of molecular-level monitoring. The performance of the proposed method, using only a limited set of informative genes, is demonstrated to be comparable or better than results reported in the literature. Furthermore, the use of the proposed method for predicting patient treatment outcome was demonstrated on the microarray data sets. Thus, the proposed methods can potentially be used to guide the design of new, more effective approaches for cancer treatment.

Acknowledgements

The authors kindly thank the anonymous reviewers for valuable comments.

Nomenclatures

$\mathbf{b}$	=	bias
$d_{ij}^{(w)}$	=	weighted Euclidean distance between i and j
$\mathbf{D}$	=	symmetric matrix of the weighted distance
DPLS	=	discriminant partial least squares
K	=	the desired number of the selected genes
m	=	the number of training samples
p	=	cancer class number
q_i	=	number of observation for class i
$\mathbf{S}$	=	covariance matrix
$\mathbf{S_B}$	=	between-class scatter matrix
$\mathbf{S_W}$	=	within-class scatter matrix
VIP	=	variable importance in the projection
$\mathbf{w}$	=	weight vector
w_{ak}	=	PLS weight for the gene expression profiles
$\mathbf{X}$	=	gene expression data matrix
$\mathbf{x}_i$	=	input vector
	=	mean vector
$\mathbf{Y}$	=	response variables (cancer labels)
$\mathbf{y}_i$	=	output label

Literature Cited

Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt; "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature*, **403**, 503–511 (2000)

Alon, U., N. Barkai, D. A. Notterman, K. Gish, Y. Barra, D. Mach and A. J. Levine; "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proc. Natl. Acad. Sci*., **96**, 6745–6750 (1999)

Alter, O., P. O. Brown and D. Botstein; "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proc. Natl. Acad. Sci*., **97**, 10101–10106 (2001)

Antonov, A. V., I. V. Tetko, M. T. Mader, J. Budczies and H. W. Mewes; "Optimization Models for Cancer Classification: Extracting Gene Interaction Information from Microarray Expression Data," *Bioinformatics*, **20**, 644–652 (2004)

Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker and M. Meyerson; "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proc. Natl. Acad. Sci*., **98**, 13790–13795 (2001)

Bicciato, S., M. Pandin, G. Didone and C. Di Bello; "Pattern Identification and Classification in Gene Expression Data Using an

Autoassociative Neural Network Model," *Biotechnol. Bioeng.*, **81**, 594–606 (2002)
Brazma, A. and J. Vilo; "Gene Expression Data Analysis," *FEBS Lett.*, **480**, 17–24 (2000)
Brown, M. P. S., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. M. Ares and D. Haussler; "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *Proc. Natl. Acad. Sci.*, **97**, 262–267 (2000)
Bullinger, L., F. G. Rucker, S. Kurz, J. Du, C. Scholl, S. Sander, A. Corbacioglu, C. Lottaz, J. Krauter and S. Frohling; "Gene-Expression Profiling Identifies Distinct Subclasses of Core Binding Factor Acute Myeloid Leukemia," *Blood*, **110**, 1291–1300 (2007)
Cho, J., D. K. Lee, J. H. Park, K. W. Kim and I. Lee; "Optimal Approach for Classification of Acute Leukemia Subtypes Based on Gene Expression Data," *Biotechnol. Prog.*, **18**, 847–854 (2002)
Cho, J., D. K. Lee, J. H. Park and I. Lee; "New Gene Selection Method for Classification of Cancer Subtypes Considering Within-Class Variation," *FEBS Lett.*, **551**, 3–7 (2003)
Cho, J., D. K. Lee, J. H. Park and I. Lee; "Gene Selection and Classification from Microarray Data Using Kernel Machine," *FEBS Lett.*, **571**, 93–98 (2004)
Chow, M. L., E. J. Moler and I. S. Mian; "Identifying Marker Genes in Transcription Profiling Data Using a Mixture of Feature Relevance Experts," *Physiol. Genomics*, **5**, 99–111 (2001)
Cremers, D., T. Kohlberger and C. Schnorr; "Shape Statistics in Kernel Space for Variational Image Segmentation," *Pattern Recognition*, **36**, 1929–1943 (2003)
Cristianini, N. and J. Shawe-Taylor; An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, U.K. (2000)
Dettling, M. and P. Buhlmann; "Supervised Clustering of Genes," *Gen. Biology*, **12**, 0069.1–0069.15 (2002)
Dorrie, J., H. Gerauer, Y. Wachter and S. J. Zunino; "Resveratrol Induces Extensive Apoptosis by Depolarizing Mitochondrial Membranes and Activating Caspase-9 in Acute Lymphoblastic Leukemia Cells," *Cancer Res.*, **61**, 4731-4739 (2001)
Duda, R. O., P. E. Hart and D. G. Stork; Pattern Classification, 2nd ed., John Wiley & Sons, New York, U.S.A. (2001)
Dudoit, S., J. Fridlyand and T. P. Speed; "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Stat. Assoc.*, **97**, 77–87 (2002)
Eriksson, L., J. L. M. Hermens, E. Johansson, H. J. M. Verhaar and S. Wold; "Multivariate Analysis of Aquatic Toxicity Data with PLS," *Aquat. Sci.*, **57**, 1015–1621 (1995)
Furey, T. S., N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler; "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," *Bioinformatics*, **16**, 906–914 (2000)
Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasen, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing and M. A. Caligiuri; "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, **286**, 531–537 (1999)
Hampton, G. M. and H. F. Frierson; "Classifying Human Cancers by Gene Expression Analysis," *Trends Mol. Med.*, **9**, 5–10 (2003)
Hastie, T., R. Tibshirani and J. Friedman; The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, U.K. (2001)
Haykin, S.; Neural Networks: A Comprehensive Foundation, Prentice Hall, Upper Saddle River, U.S.A. (1999)
Hedenfalk, I., D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg and J. Trent; "Gene-Expression Profiles in Hereditary Breast Cancer," *New Engl. J. Med.*, **344**, 539–548 (2001)
Hwang, D. H., W. A. Schmitt, G. Stephanopoulos and G. Stephanopoulos; "Determination of Minimum Sample Size and Discriminatory Expression Patterns in Microarray Data," *Bioinformatics*, **18**, 1184–1193 (2002)
Ishida, S., Y. Shigemoto-Mogami, H. Kagechika, K. Shudo, S. Ozawa, J. Sawada, Y. Ohno and K. Inoue; "Clinically Potential Subclasses of Retinoid Synergists Revealed by Gene Expression Profiling," *Molecular Cancer Therapeutics*, **2**, 49–58 (2003)
Johnson, R. A. and D. W. Wichern; Applied Multivariate Statistical Analysis, Prentice Hall, Englewood Cliffs, U.S.A. (1992)
Kroon, E., J. Krosl, U. Thorsteinsdottir, S. Baban, A. M. Buchberg and G. Sauvageau; "HoxA9 Transforms Primary Bone Marrow Cells through Specific Collaboration with Meis1a but Not Pbx1b," *EMBO J.*, **17**, 3714–3725 (1998)
Kulkarni, A., V. K. Jayaraman and B. D. Kulkarni; "Knowledge Incorporated Support Vector Machines to Detect Faults in Tennessee Eastman Process," *C&C Eng.*, **29**, 2128–2133 (2005)
Landgrebe, J., W. Wurst and G. Welzl; "Permutation-Validated Principal Components Analysis of Microarray Data," *Genome Biol.*, **3**, 0019.1–0019.11 (2002)
Lee, K. E., N. Sha, E. R. Dougherty, M. Vannucci and B. K. Mallick; "Gene Selection: a Bayesian Variable Selection Approach," *Bioinformatics*, **19**, 90–97 (2003)
Li, J. and L. Wong; "Identifying Good Diagnostic Gene Groups from Gene Expression Profiles Using the Concept of Emerging Patterns," *Bioinformatics*, **18**, 725–734 (2002)
Li, Y., C. Campbell and M. Tipping; "Bayesian Automatic Relevance Determination Algorithms for Classifying Gene Expression Data," *Bioinformatics*, **18**, 1332–1339 (2002)
Liang, J. and S. Kachalo; "Computational Analysis of Microarray Gene Expression Profiles: Clustering, Classification and Beyond," *Chem. Int. Lab. Sys.*, **62**, 199–213 (2002)
Lu, Y. and J. Han; "Cancer Classification Using Gene Expression Data," *Inf. Sys.*, **28**, 243–268 (2003)
Lyons-Weiler, J., S. Patel and S. Bhattacharya; "A Classification-Based Machine Learning Approach for the Analysis of Genome-wide Expression Data," *Genome Res.*, **13**, 503–512 (2003)
Méndez, M. A., C. Hödar, C. Vulpe, M. González and V. Cambiazo; "Discriminant Analysis to Evaluate Clustering of Gene Expression Data," *FEBS Lett.*, **522**, 24–28 (2002)
Moler, E. J., M. L. Chow and I. S. Mian; "Analysis of Molecular Profile Data Using Generative and Discriminative Methods," *Physiol. Genomics*, **4**, 109–126 (2000)
Nguyen, D. V. and D. M. Rocke; "Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data," *Bioinformatics*, **18**, 39–50 (2002a)
Nguyen, D. V. and D. M. Rocke; "Multi-Class Cancer Classification via Partial Least Squares with Gene Expression Profiles," *Bioinformatics*, **18**, 1216–1226 (2002b)
Quackenbush, J.; "Computational Analysis of Microarray Data," *Nat. Rev. Genet.*, **2**, 418–427 (2001)
Questier, F., R. Put, D. Coomans, B. Walczak and Y. vander Heyden; "The Use of CART and Multivariate Regression Trees for Supervised and Unsupervised Feature Selection," *Chem. Intel. Lab. Sys.*, **76**, 45–54 (2005)
Schölkopf, B.; Statistical Learning and Kernel Methods, Technical Report (MSR-TR-2000-23), Microsoft Research, Cambridge, U.K. (2000)
Shipp, M. A., K. N. Ross, P. Tomayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, D. S. Neuberg, E. S. Lander, J. C. Aster and T. R. Golub; "Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning," *Nat. Med.*, **8**, 68–74 (2002)
Stephanopoulos, G., D. H. Hwang, W. A. Schmitt, J. Misra and G. Stephanopoulos; "Mapping Physiological States from Microarray Expression Measurements," *Bioinformatics*, **18**, 1054–1063 (2002)

Sun, H.; "A Universal Molecular Descriptor System for Prediction of LogP, LogS, LogBB, and Absorption," *J. Chem. Inf. Comput. Sci.*, **44**, 748–757 (2004a)

Sun, H.; "Prediction of Chemical Carcinogenicity from Molecular Structure," *J. Chem. Inf. Comput. Sci.*, **44**, 1506–1514 (2004b)

Takahashi, H. and H. Honda; "A New Reliable Cancer Diagnosis Method Using Boosted Fuzzy Classifier with a SWEEP Operator Method," *J. Chem. Eng. Japan*, **38**, 763–773 (2005)

Takahashi, H., S. Tomida, T. Kobayashi and H. Honda.; "Inference of Common Genetic Network Using Fuzzy Adaptive Resonance Theory Associated Matrix Method," *J. Biosci. Bioeng.*, **96**, 154–160 (2003)

Takahashi, H., K. Masuda, T. Ando, T. Kobayashi and H. Honda; "Prognostic Predictor with Multiple Fuzzy Neural Models Using Expression Profiles from DNA Microarray for Metastases of Breast Cancer," *J. Biosci. Bioeng.*, **98**, 193–199 (2004)

Takahashi, H., T. Kobayashi and H. Honda; "Construction of Robust Prognostic Predictors by Using Projective Adaptive Resonance Theory as a Gene Filtering Method," *Bioinformatics*, **21**, 179–186 (2005)

Thomas, J. G., J. M. Olson, S. J. Tapscott and L. P. Zhao; "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles," *Genome Res.*, **11**, 1227–1236 (2001)

Tibshirani, R., T. Hastie, B. Narasimhan and G. Chu; "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Natl. Acad. Sci.*, **99**, 6567–6572 (2002)

Vapnik, V.; The Nature of Statistical Learning Theory, Springer-Verlag, New York, U.S.A. (1995)

Wang, X., Y. Wang and L. Wang; "Improving Fuzzy c-Means Clustering Based on Feature-Weight Learning," *Pattern Recognit. Lett.*, **25**, 1123–1132 (2004)

Wang, Y., I. V. Tetko, M. A. Hall, E. Frank, A. Facius, K. F. X. Mayer and H. W. Mewes; "Gene Selection from Microarray Data for Cancer Classification—a Machine Learning Approach," *Comput. Biol. Chem.*, **29**, 37–46 (2005)

Xiong, M. M., W. Li, J. Zhao, L. Jin and E. Boerwinkle; "Feature (Gene) Selection in Gene Expression-Based Tumor Classification," *Mol. Genet. Metabol.*, **73**, 239–247 (2001)

Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai and T. P. Speed; "Normalization for cDNA Microarray Data: a Robust Composite Method Addressing Single and Multiple Slide Systematic Variation," *Nucl. Acid. Res.*, **30**, e15 (2002)

Yeung, K. Y. and W. L. Ruzzo; "Principal Component Analysis for Clustering Gene Expression Data," *Bioinformatics*, **17**, 763–774 (2001)

Yoo, C. K., I. Lee and P. A. Vanrolleghem; "Interpreting Patterns and Analysis of Acute Leukemia Gene Expression Data by Multivariate Fuzzy Statistical Analysis," *Comput. Chem. Eng.*, **29**, 1345–1356 (2005)

Zhang, H. P., C. Y. Yu, B. T. Singer and M. M. Xiong; "Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data," *Proc. Natl. Acad. Sci.*, **98**, 6730–6735 (2001)