Genomics & Informatics Vol. 2(1) 19-29, March 2004

Informative Gene Selection Method in Tumor Classification

Hyosoo Lee and Jong Hoon Park*

Department of Biological Science, Sookmyung Women's University, Seoul 140-742, Korea

Abstract

Gene expression profiles may offer more information than morphology and provide an alternative to morphology- based tumor classification systems. Informative gene selection is finding gene subsets that are able to discriminate between tumor types, and may have clear biological interpretation. Gene selection is a fundamental issue in gene expression based tumor classification. In this report, techniques for selecting informative genes are illustrated and supervised shaving introduced as a gene selection method in the place of a clustering algorithm. The supervised shaving method showed good performance in gene selection and classification, even though it is a clustering algorithm. Almost selected genes are related to leukemia disease. The expression profiles of 3051 genes were analyzed in 27 acute lymphoblastic leukemia and 11 myeloid leukemia samples. Through these examples, the supervised shaving method has been shown to produce biologically significant genes of more than 94% accuracy of classification. In this report, SVM has also been shown to be a practicable method for gene expression-based classification.

Keywords: gene expression; gene selection; gene shaving; microarray; tumor classification

Introduction

DNA microarray technology generated a panoramic survey of genes expressed in a sample of cells. When the samples correspond to different pathological states of the same tissue or subtypes of the sample malignancy, transcription profiling holds promise as a method for classifying and analyzing cancer from a molecular rather than morphological perspective (Alizadeh *et al.*, 2000, Alon *et al.*, 1999). Several studies have used arrays to analyze gene expression in the colon, breast and other tumors, and these studies have demonstrated the potential utility of expression profiling for classifying tumors (Perou *et al.*, 1999).

Informative Gene selection is an important component for gene expression-based tumor classification systems. A large numbers of genes increase the computational complexity and cost, and compromise the generalization properties of the classifier. It is recognized that the higher the ratio of the number of training samples to the number of free classifier parameters, the better the generalization properties of the resulting classifier. A large number of genes will improve the estimation of the classification error. Therefore, reducing the dimensionality of the gene expression information is a key issue in developing a successful gene expression-based tumor classification system. In addition to reducing noise and improving the accuracy of tumor classification, a selected subset of genes, with high accuracy of classification, may be involved in the pathways or some biological process leading to tumor development and have important biological meaning. That is, these informative genes represent putative targets for therapeutic agents and understanding the basic biology of the disorder. A typical profiling study measures the expression levels of thousands of genes (features) L across tens of samples N, with each samples labeled as being of one type or another. The problem to be considered here is that of identifying marker genes given N labeled L-features samples.

The method introduced by Golub *et al.* (1999) was examined to focus the selection of genes that appear to be the best diagnostic indicators. This amounts to a kind of dimensionality reduction of the dataset, and shows good performance in tumor type classification. However, the subset genes were found not to correlate with each other, and some genes do not have a biological meaningful explanation. In order to improve this gene selection method, additional effort is needed to develop ways of identifying meaningful genes in these types of dataset. (Furey *et al.*, 2000)

In this report, we propose 'supervised shaving' clustering analysis of gene expression data is proposed for the identification of gene subsets. The 'supervised shaving' is designed to extract coherent and small clusters of genes that vary as much as possible across the

* Corresponding author: E-mail parkjh@sookmyung.ac.kr, Tel +82-2-710-9414, Fax +82-2-2077-7322
Accepted 11 March 2004

samples. In addition, 'supervised shaving' is proposed via the incorporation of other prognostic factors in the search for interesting gene clusters (Hastie *et al.,* 2000). As a result of 'supervised shaving' as a gene selection method, more biologically meaningful subsets of genes were obtained in a cluster.

The gene selection issue has been addressed under a classification framework that may be more relevant to the clinical application in a diagnosis. To this end, the use of SVM has been present, as a supervised machine learning technique, for the classification algorithms. SVMs have been shown to perform well in multiple areas of biological analysis, including evaluating microarray expression data (Brown *et al.*, 2000). SVMs have demonstrated the ability to not only correctly separate entities into appropriate classes, but also identify instances whose established classification is not supported by the data. Expression datasets contain measurements for thousands of genes, which prove problematic for many traditional methods. SVMs are well suited to working with high dimensional data such as this.

To test the generality of the approach, experiments were run using the leukemia data from Golub *et al.* (1999) (72 patient samples). Clustered genes, due to PCA based 'supervised shaving' were shown to achieve high accuracy of classification, and that the selected genes can be used as biomarkers for tumor classification.

Methods

Gene expression data from acute leukemia

A data set, well known for human acute leukemia samples, originally analyzed by Golub *et al.* was used. The raw dataset can be obtained at: http://www.broad.mit.edu/cgi-bin/cancer/ datasets.cgi. The dataset consisted of the expression profiles of 6817 genes from 47 ALL and 25 AML samples.

The first step in out process of analyzing these data was to perform the basic transformations reported in Golub 1999. The transformations set the minimum expression values to 100 and the maximum to 16000, and then the genes that are beyond these limits are filtered out. The

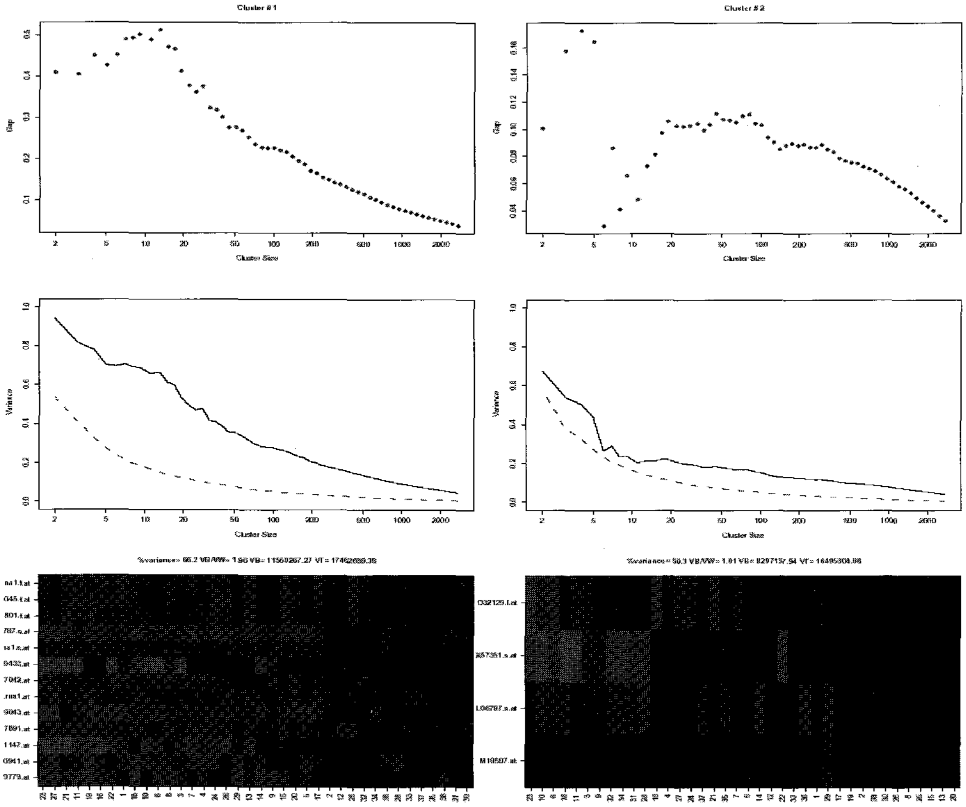

Fig. 1. Supervised shaving result. The left column represent cluster 1 and the right cluster 2. The first row is the gap curve estimated cluster size and the second the variance plots for real and randomized data. The third row is the gene expression level across the samples in each cluster. Almost all the left side (green) samples are ALL and those on the right side (red) are AML.

filtering process selected 3051 genes that seemed worthy of further investigation. The filtering was not especially robust, in the sense that it was based on the minimum or maximum expression values between subjects. This was a non-specific filter. The genes were selected according to their variability, and not with respect to their ability to classify any particular set of samples. In this analysis, the genes were selected according to their behavior in the training set. If the same gene selection methods were applied to the test set, a different set of genes would be selected in the test set, despite the identical selection process. Due to the need to compare and combine the analyses that approach could not be used: the same genes, not just the same filters, must be employed in both phases of the analysis.

Multiple t-tests

Permutation adjusted p-values for the maxT procedure, with Welch t-statistics, were computed. The adjusted p-values for the maxT and minP step-down multiple testing procedures are described in Westfall and Young (1993). The bioconductor's multitest packages were used, which contain a collection of functions for multiple hypotheses testing, implemented using R language. These functions can be used to identify differentially expressed genes in microarray experiments, that is, genes whose expression levels are associated with a response or covariate of interest. The 17 genes with the smallest adjusted p-value were selected for the classification analysis.

Prediction strength (PS) and gene selection

Golub *et al.* proposed the use of a collection of known samples to generate a "class predictor" which is then able to assign a new sample to one of two classes. This predictor is created with the aid of 'prediction strength'. Let $[\mu_1(g), s_1(g)]$ and $[\mu_2(g), s_2(g)]$ denote the means and standard deviation of the log of the expression levels of gene g in two samples, respectively. Then, PS is defined as

$$PS(g) = \left| \frac{\mu_1(g) - \mu_2(g)}{s_1(g) - s_2(g)} \right|$$

We calculate PS scores, which give the highest score to those genes whose expression levels, differ most on average in the two classes while also favoring those with small deviations in scores in the respective classes. We then simply take the genes with the top 17 ranked PS scores as our top features.

'Supervised shaving' as a gene selection method

Gene shaving, introduced by Hastie *et al.*, is a method for exploratory analysis of gene expression microarray data. It tries to identify subset of genes with coherent expression patterns with large variation across samples. The method differ form hierarchical clustering method in that genes may belong to more than one cluster, and the clustering can be un-supervised, treat the genes and samples as unlabeled, or partially or fully supervised by known properties of genes/samples.

Gene shaving is an iterative algorithm based on the principal components or the singular value decomposition (SVD) of the data matrix. It starts with the entire microarray gene expression matrix A and seeks a function of the genes in the direction of maximal variation across the samples. The simplest form of this function is a normalized linear combination of the genes weighted by its largest principal component loadings, referred to as the super gene. The genes may be sorted according to the principal component weights. A fraction α of the genes having lowest correlation (essentially the absolute inner product) with the super gene are then shaved off from the original data matrix. The process of calculating the leading principal component and shaving off some genes is iterated on the reduced data matrix until only two genes remain. This iterative to-down process produces a sequence of nested gene blocks of sizes ranging from the full set of N genes down to the final block consisting of just two genes.

The method requires a quality measure for a cluster. In particular, a Gap estimation is used to select the optimal number of genes in a cluster, which is based on the between and within variances of the clusters form the raw data matrix and its permutation. The number of permutations should be specified in the analysis of gene shaving for a microrray data. The next step is to remove the effect of genes in the optimal cluster, C_1 say, from the original matrix A. By computing the average gene or the vector of column average for C_1, denoted by C_1, we can remove the component that is correlated with this average. This is equivalent to regressing each row of A on the average gene row C_1, and replacing the former with the regression residuals. Such a process is referred to as orthogonalization by Hastie *et al.*, from which a modified data matrix A_{ortho} is produced. With A_{ortho}, the whole process is repeated of calculating the leading principal component, producing another nested sequence of shaved gene blocks, applying the Gap statistic to obtain the next optimal cluster C_2, and orthogonalizing the current data matrix. This sequence of operations is iterated until L gene clusters C_1, Λ, C_L

are found, which can be displayed graphically for visual inspection.

'Supervised shaving' is a modification of gene shaving. Full or partial supervised shaving for class discrimination can be carried out if the information of the column (sample) classification is available. In particular, an indicator vector with the length of samples is needed to specify the sample classification.

When fully supervised, the shaving procedure reduces to simply ranking the genes. Thus there is no role for principal components or top-down shaving in this case. However, to encourage coherence within the gene clusters, it can be better to use a partially supervised criterion, which does use principal components and top-down shaving.

In case of Golub's acute leukemia data, we use labeled 38 training data sets samples with ALL or AML and partially supervised shaving method with 10 times permutation.

SVMs training and evaluation.

SVMs are a relatively new type of learning algorithm, originally introduced by Vapnik and co-workers (Vapnik, 1998) and successively extended by a number of other researchers. Their remarkably robust performance with respect to sparse and noisy data is making them the system of choice in a number of applications from text categorization to protein function prediction.

When used for classification, they separate a given set of binary-labeled training data with a hyper-plane that is maximally distant from them (known as 'the maximal margin hyper-plane'). For cases in which no linear separation is possible, they can work in combination with the technique of 'kernels'. That automatically realizes q non-linear mapping to a feature space. The hyper-plane found by the SVM in feature space corresponds to a non-linear decision boundary in the input space. That is, estimating an SVM requires specifying an inner-product kernel function, a measure of similarity between two profile vectors $X^i_L = \{x^i_1, K, x^i_L\}$ and $X^j_L = \{x^j_1, K, x^j_L\}$, where x^i_L and x^j_L are gene expression level of gene L for samples belonging to class i and j. Since there is no general theory for determining the most appropriate kernel for a particular learning problem, 'radial basis' kernel function suggested that superior to the other methods (Brown *et al.*, 1999). 'Radial basis' kernel function $K(X^i_L, X^i_L) = \exp(- \| X^i_L - X^i_L \|^{2*} \gamma)$, where $\gamma = 1/2\sigma^2$ is a user-defined width parameter.

Because of the limited number of training examples, a leave-one-out cross validation strategy was utilized. A pool of N training data set was partitioned into two disjoint sets. The estimation set, N-1 examples, was used to determine the parameters of an SVM, and the test set, 1 example, was used to assess its generalization performance (here, N=38). The label assigned by a trained SVM to a test example can be a true positive (known positive test example, assigned positive label), true negative (negative example, negative label), false positive (negative example, positive label), or false negative (positive example, negative label).

Golub's 72 leukemia samples were partitioned into estimation and test set containing 38 and 34 samples, respectively. The generalization performance of this '38 estimation, 34 test' partitioning is how many of the 34 test examples were assigned to be true positives or true negatives.

Results

Gene selection method comparison

In order to compare the 'supervised shaving' method with other gene selection techniques, other methods; multiple t-test and 'Prediction Strength' (PS method) were also addressed. Both of these have previously been applied to the gene selection method in another published study (Xiong *et al.*, 2001). The gene list in table 1-3 represents the result of each gene selection method. 17 genes in two clusters from 'supervised shaving'(Fig. 1), and 17 top ranked genes in two other methods were obtained. To ensure fairness in the gene numbers, the same numbers of genes from three methods were obtained. Although the precise composition of the 17 genes differed, each set of genes was effective in terms of discriminating between different tumor types. The small overlap in terms of the specific genes may suggest the presence of many gene subsets for a given cardinality that are equally well generalized. However, there are no common genes in the multiple tests; the PS and supervised shaving methods. Only small overlaps existed between the two methods. It should be understood, that these three methods have very different basic concept in analysis. In other words, PCA based 'supervised clustering' captures the variance in a dataset in terms of principle components, but other two methods are not dependent on variation but difference of mean values of each samples. A union set of three gene lists was also made, and trained with a new combined data matrix for leukemia cancer classification. However, the result of classification was not as good as expected. (Data are not shown)

Experimental studies of informative genes as ALL and AML markers

To understand the scientific and clinical relevance of our results, a PubMed keyword search was conducted using gene name words, such as 'leukemia' and 'acute leukemia'. The most important genes found have also been examined in the related contents.

Table 1 shows the gene list selected by multiple t-tests. TCF3 (M31523), E2A immunoglobulin enhancer binding factors E12/E47, is associated with childhood B cell acute lymphoblastic leukemia. The resulting fusion of the 5' *E2A* sequences with the 3' portions of other genes leads to the expression of two well-characterized fusion proteins, E2A-PBX1 and E2A-HLF. Since the E2A, PBX1 and HLF proteins appears to function as transcription factors, it appears likely that the oncogenic fusions appear to function in development by causing abnormal transcriptional regulation of key target genes (LeBrun *et al.*, 2003). Fig. 2 (a) shows that *TCF3* is differentially expressed in a training sample (p=0.013). For the genes in Table 2, Parisi *et al.* explored new therapeutic approaches to AML focused on immune-based therapy through monoclonal antibodies that target and destroy malignant cells via specific cell receptors (Parisi *et al.*, 2002). One such agent is gemtuzumab (CAN-676), an agent that targets the *CD33* (M23197) antigen on malignant myeloid cells, and initial studies have shown significant anticancer activity. In another study (Hamann *et al.*, 2002), it was reported that *CD33* is expressed by AML cell in 〉80% of patients, but not by normal hematipoietic stem cells, suggesting that eliminating CD33(+) cells may be

Table 1. Subset of genes using multiple t statistics

Accession	Symblo	Gene Annotation
D26156	HLA-A	Transcriptional activator hSNF2b
L47738	CYFIP2	Inducible protein mRNA
M11147	FTL	FTL Ferritin, light polypeptide
M31211	MLC1SA	MYL 1 Myosin light chain (alkali)
M31523	TCF3	TCF3 Transcription factor 3 (E2A immunoglobuling enhancer binding factors E12/E47)
M55150	FAH	FAH Fumarylacetoacetate
M92287	CCND3	CCND 3 Cycling D3
S50223	ZNF22	HKR-T1
U05259	CD79A	MB-1 gene
U22376	MYB	C-myb gene extracted from Human (c-myb) gene, complete primary CDs, and five complete alternatively spliced CDs
U50136	LTC4S	Leukotriene C4 synthase (LTC 4S) gene
U82759	HOXA9	GB DEF = Homeodomain protein HoxA9 mRNA
X59417	PSMA6	PROTEASOME IOTA CHAIN
X63469	GTF2E2	GTF2E2 General transcription factor TFIIE beta subunit, 34kD
X74262	RBBP4	RETINOBLASTOMA BINDING PROTEIN P48
X95735	ZYX	Zyxin
Z15115	TOP2B	TOP2B Topoisomerase (DNA) II beta (180kD)

Table 2. Subset of genes using Prediction Strength (PS) method

Accession	Symblo	Gene Annotation
M16038	LYN	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
M23197	CD33	CD33 CD33 antigen (differentiation antigen)
M27891	CST3	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
M55150	FAH	FAH Fumarylacetoacetate
M84526	DF	DF D component of complement (adipsin)
M92287	CCND3	CCND3 Cyclin D3
M90020	AZU1	Azurocidin gene
U05259	CD79A	MB-1 gene
U22376	MYB	C-myb gene extracted from Human (c-myb) gene, complete primary CDs, and five complete alternatively spliced CDs
U46751	SQSTM1	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
U50136	LTC4S	Leukotriene C4 synthase (LTC4S) gene
U82759	HOXA9	GB DEF = Homeodomain protein HoxA9 mRNA
X17042	PRG1	PRG1 Proteoglycan 1, secretory granule
X59417	PSMA6	PROTEASOME IOTA CHAIN
X95735	ZYX	Zyxin
Y00787	IL8	INTERLEUKIN-8 PRECURSOR
Y12670	OBRGRP	LEPR Leptin receptor

therapeutically beneficial. The 'leptin receptor gene – related protein' (Y12670) was also found in the 'PS method' gene list (Table 2). Recently, leptin has been shown to play a regulatory role for differentiation within the myeloid and erythroid cell lineage. The research on whether leptin plays a role in ALL shows that the leptin levels in bone marrow–derived plasma of children with ALL were significantly lower than in those of healthy controls (Wex *et al.*, 2002).

CCND3 (M92287), *CD79A* (U05259), *MYB* (U22376), *HOXA9* (U82759) and *ZYX* (X95735) are common in two gene lists, but are not in the 'supervised shaving' list. For those genes in the two common methods, Sonoki *et al.* found that cyclin D3 (*CCND3*) was a target gene of mature B cell malignancies. (Sonoki *et al.*, 2001). Another study (Krissansen *et al.*, 1986) concluded that glucocorticoids cause G_0/G_1 arrest of lymphoid cells and, at least in part, to a decrease in the abundance of the G_1 progression factor, cyclin D3, which manages regulation of the cell cycle. The mRNA encoding cyclin D3 (CcnD3 mRNA) was rapidly down regulated when dexamethasone was added to P1798 murine T lymphoma cells.

CD79A is a cytoplasmic antigen that acts as a mediator of signal transduction from the cell surface to the cytoplasm in association with CD79b, and is expressed early in B–cell development (Lai *et al.*, 2000). Frater et al. reported that *CD79a* has been so clearly associated with acute lymphoblastic leukemia (ALL) by some researchers that its expression in the presence of blast markers is considered indicative of B–ALL (Frater *et al.*, 2003). From Fig. 2(b) this report could be true.

c–myb is essential for the development of hematopoietic cells, and plays a role in proliferation, anti–apoptosis and differentiation. *c–myb* can be activated to transform myeloid cells, either by over expression at the transcriptional level or protein truncation, which removes sequences critical for phosphorylation–induced ubiquitinization and 26S–proteasome degradation (Bies *et al.*, 1997, 1999).

Cytogenetic, genetic and functional studies have demonstrated a direct link between deregulated *Hoxa9* expression and acute myeloid leukemia (AML). The leukemogenic potential of *Hoxa9* was directly demonstrated by the development of AML in mouse bone marrow transplantation chimeras that received a graft

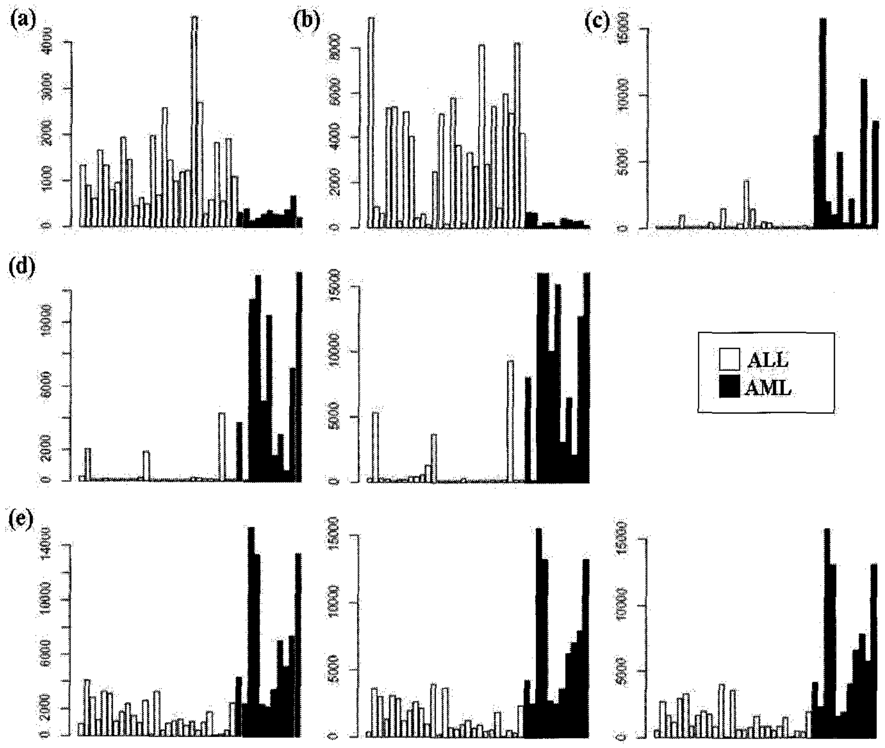

Fig. 2. Gene expression level across the acute leukemia samples; 38 training sets, 27 ALL (white bar) and 11 AML (black bar). The first row of panels (a), (b) and (c) correspond to the *TCF3*, *CD79A* and *MPO* genes. Panel (d), the second row of the two bar–plots represents the *IL8* transcripts, M28130 and Y00787. Panel (e), the last row of the three bare plots corresponds to lysozyme mRNA, J03801, M19045 and X14008.

of primitive hematopoietic cells engineered by retroviral gene transfer to over express *Hoxa9* (Kroon *et al.*, 1998).

In Table 3 the biologically important genes obtained from the 'supervised shaving' clustering method are also shown. The enzyme myeloperoxidase (M19507) is a well-established marker of myeloid differentiation. Most myeloid leukemia expresses *MPO* enzyme activity at the light microscopic level, whereas lymphoid leukemia characteristically lacks such expression. However, the diagnostic significance of *MPO* RNA or immunohistochemically detectable *MPO* protein expressions in leukemic blasts is unclear. It has been demonstrated that *MPO* expression shows no significant correlation with other markers of myeloid differentiation (Austin *et al.*, 1998). However, Fig. 2(c) shows MPO was over expressed in acute myeloid leukemia.

The chemokine receptor *CXCR4* (L06797) is important on acute lymphoblastic leukemia, which is based on the research about stromal cell-derived factor -1(*SDF-1*) inhibitor (Juarez *et al.*, 2003). *SDF-1* is a key regulator of the behavior of normal and leukemic precursor-B (pre-B) cells. It is possible that inhibiting *SDF-1* driven processes in pre-B acute lymphoblastic leukemia (ALL) may have therapeutic implications. *SDF-1* and *CXCR4* have been implicated in numerous disease state, including cancer metastasis, cell infiltration into arthritic joints and HIV infection, as well as survival and dissemination of leukemic blasts.

In addition, there are some remarkable points in the gene list shown in Table 3. Some transcripts derived from same gene were found in the list. M28130 and Y00787 are mRNA transcripted from the *IL8* (Interleukin 8) gene, which are detected in peripheral leukemic cells obtained from adult T-cell leukemia patients, as well as in cultured human T-cell leukemia virus type 1-infected T cell lines. Detection of *IL-8* and *IL-8R* might help to identify ALL types, predict prognosis and the development of CNSL, which is established by Liu *et al.*, 1999. Another example is lysozyme, which has J03801, M19045 and X14008 for transcripts. The leukemic lymphoblasts were characterized by low levels of lysozyme, as compared to by the leukemic myeloblasts or normal lymphocytes (Ho *et al.*, 1984). It is worth noting, *IL8* or Lysozyme transcripts showed similar expression patterns across the leukemia samples (Fig. 2). It means that this transcript expression is reliable and samples are subdivided. Such a phenomenon was not found in the other gene list table.

The ferritin related genes, *FTH1* (L20941) and *FTL* (M11147), are shown in Table 3. High serum ferritin levels, without any correspondence to the amount of total body iron storage, have been found in patients with leukemia. Investigating 96 adults with different types of leukemia, serum ferritin can be used as a tumor marker in myeloid. So, the serum ferritin concentration must be valued as a clinically useful tumor marker in these types of leukemia, exhibiting a helpful and simple parameter in monitoring the activity of the disease (Aulbert *et al.*, 1985).

Classification using SVM (Support Vector Machine)

As mentioned above, bone marrow or peripheral blood samples from 72 patients with other acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) were used in this study. The data was split into a training set consisting of 38 samples, of which 27 were ALL and

Table 3. Subset of genes using supervised shaving

Accession	Symblo	Gene Annotation
L19779	HIST2H2AA	Histone H2A. 2 mRNA
L20941	FTH1	FTH1 Ferritin heavy chain
M11147	FTL	FTL Ferritin, light polypeptide
M19507	MPO	MPO Myeloperoxidase
M27891	CST3	CST3 Cystatic C (amyloid angiopathy and cerebral hemorrhage)
M69043	NFKBIA	MAJOR HISTO COMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3
M96326	AZU1	Azurocidin gene
X17042	PRG1	PRG1 Proteoglycan 1, secretory granule
Y00433	GPX1	GPX1 Glutathione peroxidase 1
L06797	CXCR4	PROBABLE G PROTEIN-COUPLED RECEPTOR LCR1 HOMOLOG
X57351	IFITM2	RPS3 Ribosomal protein S3
M28130	IL8	Interleukin 8 (IL8) gene
Y00787	IL8	INTERLEUKIN-8 PRECURSOR
J03801	LYZ	LYZ Lysozyme
M19045	LYZ	LYZ Lysozyme
X14008	LYZ	Lysozyme gene (EC 3.2.1.17)
D32129	HLA-A	HLA-A MHC class I protein HLA-A (HLA-A28, -B40, -Cw3)

11 AML, and a test set of 34 samples, 20 ALL and 14 AML. The prediction on the training set using SVM is shown in Table 4.

In Table 4, in the top-left and bottom-right are correctly predicted leukemia type numbers and in the top-right and bottom-left are the misclassified cases, such as ALL incorrectly classified into AML or the reverse. Table 4(a) is the result of prediction using filtering processed data with 3051 genes, and is the worst among the results that meet our expectation. The result accuracy was 67%. Tables 4(b), (c) and (d) represent the results of the predictions using the multiple t-test, 'supervised shaving'

Table 4. Prediction of the training set using SVM. (a) is with a raw data set without gene selection processing. (b) to (d) correspond to multiple t statistics, supervised shaving and Prediction Strength method, respectively.

(a)

SVM Prediction	ALL	AML
ALL	20	11
AML	0	3

(b)

SVM Prediction	ALL	AML
ALL	20	5
AML	0	9

(c)

SVM Prediction	ALL	AML
ALL	20	2
AML	0	12

(d)

SVM Prediction	ALL	AML
ALL	20	1
AML	0	13

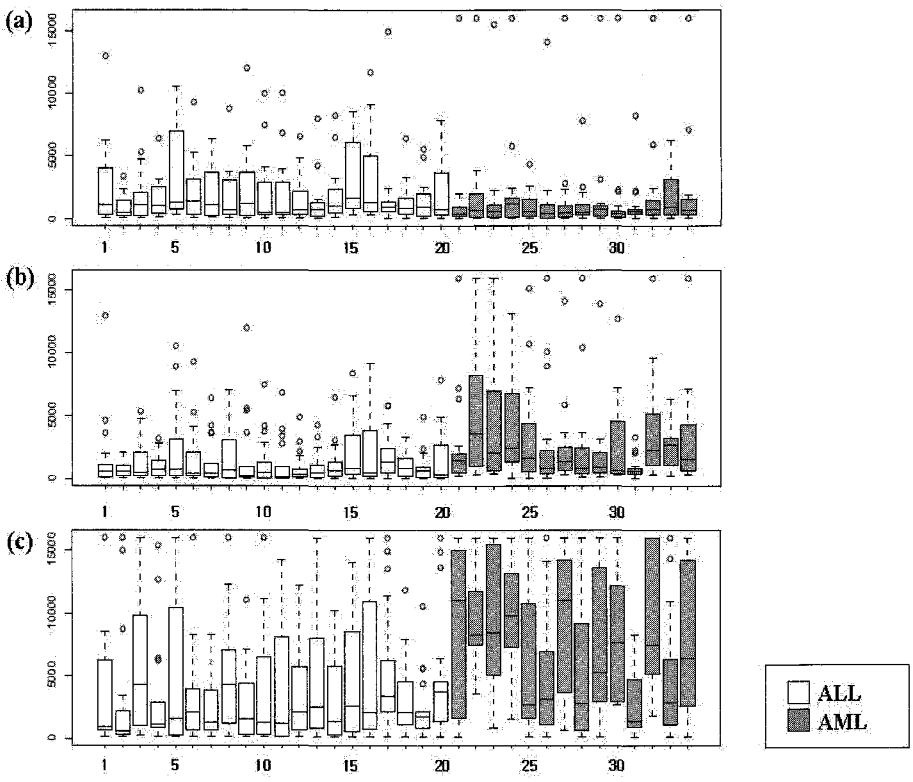

Fig. 3. Gene expression data distribution of acute leukemia samples; 38 training sets, 27 ALL (white box), 11 AML (gray box), represented by box-plot. Panel (a) represents the new data matrix with subsets of genes selected by multiple t statistics. Panels (b) and (c) are Golub' s 'Prediction Strength' and 'Supervised shaving' results, respectively.

and PS method, respectively. The prediction accuracy with 'Supervised Shaving' was 94% and with the PS method was 97%, which were better than the result of the multiple t-test, which was 85%. It was also found that the ALL type was correctly classified in all the methods, but the AML type depended on the gene selection method. AML patient number '31' in the test data was misclassified as ALL in all the SVM tests. As shown in box plots (Fig. 3), AML patient sample number '31' tended to a have low expression level compare to the other AML patients samples in all three different data representation. Therefore, we suppose that this AML sample is not general in all AML or other leukemia types, and such a sample has an effect on the misclassification and incorrect result.

Discussion

Simultaneously monitoring the expression of thousands of genes holds great promise for better understanding cancer biology and developing accurate tumor classification schemes. However, the very large amount of gene expression information provided by contemporary microarray technology causes problems for both basic research and clinical application. The high cost of large-scale microarray experiments lead to a sample size that is usually several orders of magnitude smaller than the number of genes being monitored. As a result, it is mathematically infeasible to use all the gene expression information to develop a classification algorithm for a relatively small number of tumors. It also well documented in the statistical literature that too many feature variable genes could harm the performance of the classifier. Therefore, development of an accurate tumor classification scheme must begin with selection of a subset of the initially observed genes for tumor classification.

To determine the sets of biologically meaningful genes, the available expression data from 6817 genes in 47 ALL and 25 AML samples were analyzed, and three different methods used to select subsets of genes. The PCA based 'supervised shaving' method was identified to produce subsets of genes that were more leukemia related than other methods, and the expression pattern of some genes were correlated to each other because the transcripts came from the same gene (Fig. 2(d), (e)). The 'supervised shaving' method uses information about the columns to 'supervise' the shaving of rows. In this case, leukemia type class labels were used, and supervised shaving maximizes a weighted combination of the column variance using the information. So, most genes have a higher variance between two samples than other subsets of genes. Because the 'supervised shaving' method identifies subsets of genes with coherent expression patterns and a large variance across samples, it may not find non-coherent genes, but that have a significant difference between two samples. However, most genes are very important from a biological and clinical point of view, and show good performance in classification. In the same manner, the PS method also produces high accuracy in classification, but a genes' composition is dissimilar to that of the supervised shaving gene, and some ideas may be suggested by such a different gene selection method.

First, for investigations involving a large number of observed variables, it is often useful to simplify the analysis by considering a smaller number of linear combinations of the original variable. In this case, gene expression data consist of a number of microarray experiment(samples) in thousands of genes. To identify differentially expressed genes, the multiple t-test and PS method frequently reduce the intensity from all gene expressions to a single mean value that corresponds to a sample group. This process may be inadequate to understand data representation and depends on the data itself. On the other hand, PCA based methods reduce the dimensionality of a data set, while retaining as much information as possible. So, the result may be affected by the difference of cell lines in the same sample group.

Second, deciding on the number of informative genes is a problem. All of the three compared methods have this problem. In the case of 'supervised shaving' method, the cluster number is a case in point. In this report, the cluster number was fixed as two, and corresponded to 17 genes. However, the investigator should decide the cut-off value for selecting genes in the multiple t-test and PS method. In order to do justice to the gene number, other methods also produce the 17 top ranked genes. However, only 2 or 3 genes can achieve more than 90% accuracy of classification. This result implies that a small number of selected genes may be used as biomarkers for tumor classification, or may have some relevance in tumor development, and serve as a potential drug target (Xiong *et al.*, 2001).

In conclusion, there is no general method to select the right informative gene, but the PCA based method 'supervised shaving' gives biologically and clinically meaningful genes. There is no single best procedure for selecting an optimal subset of genes. It was not the purpose of this paper to find the optimum methods for gene selection, but rather to illustrate how gene selection can be useful to cancer classification using microarray gene expression data classification. In this report, three statistical procedures were compared for their accuracy of cancer classification: multiple t tests, the PS statistic suggested by Golub *et al.* and 'supervised shaving'. The result indicates the PS method and 'supervised shaving' methods performed similarly, and both methods performed better than the t statistics. However, the 'supervised shaving' method demands much more computational time than the PS method.

An attempt to use the 'supervised shaving' clustering algorithm to select informative genes suggests the use of the clustering method as a gene selection method. The level of expressions of sets(clusters) of genes across samples correlate. As a result, it is likely that a smaller number without significant loss of information can capture the information contained in all large numbers of genes. This is a direct result of the fact that sets of genes are similarly regulated and, hence, play a similar role in cancer classification. It is known from pattern recognition theory that a good feature subset is one that contains features highly correlated with the class, yet uncorrelated with each other. In this report, it has been shown that a number of gene clusters are obtained from the 'supervised shaving' method. The expression levels of genes within the same cluster have high correlation, and the expression levels of genes in different clusters have low correlation. It is speculated that the genes within a cluster lie in a single or co-regulated pathway, and genes in different clusters lie in different pathways or in pathways that are not co-regulated. So, the progress of a clustering method to select informative coherent genes is needed for cancer classifications.

Acknowledgement

This study was supported by a grant from the International Mobile Telecommunications 2000 R&D Project, Ministry of Information and Communication, Republic of Korea (01-PJ11-PG9-01BT00A-0020).

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct 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., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96, 6745-6750.

Aulbert, E. and Schmidt, C. G. (1985). Ferritin—a tumor marker in myeloid leukemia. *Cancer Detect. Prev.* 8, 297-302.

Bies, J., Nazarov, V., and Wolff, L. (1999). Identification of protein instability determinants in the carboxy-terminal region of c-Myb removed as a result of retroviral integration in murine monocytic leukemias. *J. Virol.* 73, 2038-2044.

Bies, J. and Wolff, L. (1997). Oncogenic activation of c-Myb by carboxyl-terminal truncation leads to decreased proteolysis by the ubiquitin-26S proteasome pathway. *Oncogene* 14, 203-212.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA.* 97, 262-267.,

Frater, J. L., Yaseen, N. R., Peterson, L. C., Tallman, M. S., and Goolsby, C. L. (2003). Biphenotypic acute leukemia with coexpression of CD79a and markers of myeloid lineage. *Arch. Pathol. Lab Med.* 127, 356-359.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906-914.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

Hamann, P. R., Hinman, L. M., Hollander, I., Beyer, C. F., Lindh, D., Holcomb, R., Hallett, W., Tsou, H. R., Upeslacis, J., Shochat, D., Mountain, A., Flowers, D. A., and Bernstein, I. (2002). Gemtuzumab ozogamicin, a potent and selective anti-CD33 antibody-calicheamicin conjugate for treatment of acute myeloid leukemia. *Bioconjug. Chem.* 13, 47-58.

Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., and Brown, P. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 1, RESEARCH0003.

Ho, A. D., Fiehn, W., and Hunstein, W. (1984). Plasma and intracellular levels of lactate dehydrogenase, phosphohexose isomerase and lysozyme activity in acute leukemia. *Blut.* 49, 19-28.

Juarez, J., Bradstock, K. F., Gottlieb, D. J., and Bendall, L. J. (2003). Effects of inhibitors of the chemokine receptor CXCR4 on acute lymphoblastic leukemia cells in vitro. *Leukemia* 17, 1294-1300.

Kroon, E., Krosl, J., Thorsteinsdottir, U., Baban, S., Buchberg, A. M., and Sauvageau, G. (1998). Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *Embo J.* 17, 3714-3725.

Lai, R., Juco, J., Lee, S. F., Nahirniak, S., and Etches, W. S. (2000). Flow cytometric detection of CD79a expression in T-cell acute lymphoblastic leukemias. *Am. J. Clin. Pathol.* 113, 823-830.

LeBrun, D. P. (2003). E2A basic helix-loop-helix transcription factors in human leukemia. *Front. Biosci.* 8, s206-222.

Liu, J., Zeng, H., and Zhang, Y. (1999). [Study on the expression of interleukin-8 and its receptors in acute leukemia]. *Zhonghua Xue Ye Xue Za Zhi* 20, 24-26.

Parisi, E., Draznin, J., Stoopler, E., Schuster, S. J., Porter, D., and Sollecito, T. P. (2002). Acute myelogenous leukemia: advances and limitations of treatment. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* 93, 257-263.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O., and

Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* 96, 9212–9217.

Sonoki, T., Harder, L., Horsman, D. E., Karran, L., Taniguchi, I., Willis, T. G., Gesk, S., Steinemann, D., Zucca, E., Schlegelberger, B., Sole, F., Mungall, A. J., Gascoyne, R. D., Siebert, R., and Dyer, M. J. (2001). Cyclin D3 is a target gene of t(6;14)(p21.1;q32.3) of mature B-cell malignancies. *Blood* 98, 2837–2844.

Wex, H., Ponelis, E., Wex, T., Dressendorfer, R., Mittler, U., and Vorwerk, P. (2002). Plasma leptin and leptin receptor expression in childhood acute lymphoblastic leukemia. *Int. J. Hematol.* 76, 446–452.

Xiong, M., Li, W., Zhao, J., Jin, L., and Boerwinkle, E. (2001). Feature (gene) selection in gene expression-based tumor classification. *Mol. Genet. Metab.* 73, 239–247.