Machine Learning in DNA Microarray Analysis for Cancer Classification

Sung-Bae Cho and Hong-Hee Won
Dept. of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
sbcho@cs.yonsei.ac.kr cool@candy.yonsei.ac.kr

Abstract

The development of microarray technology has supplied a large volume of data to many fields. In particular, it has been applied to prediction and diagnosis of cancer, so that it expectedly helps us to exactly predict and diagnose cancer. To precisely classify cancer we have to select genes related to cancer because extracted genes from microarray have many noises. In this paper, we attempt to explore many features and classifiers using three benchmark datasets to systematically evaluate the performances of the feature selection methods and machine learning classifiers. Three benchmark datasets are Leukemia cancer dataset, Colon cancer dataset and Lymphoma cancer data set. Pearson's and Spearman's correlation coefficients, Euclidean distance, cosine coefficient, information gain, mutual information and signal to noise ratio have been used for feature selection. Multi-layer perceptron, k-nearest neighbour, support vector machine and structure adaptive self–organizing map have been used for classification. Also, we have combined the classifiers to improve the performance of classification. Experimental results show that the ensemble with several basis classifiers produces the best recognition rate on the benchmark dataset.

Keywords: Biological data mining, feature selection, classification, gene expression profile, MLP, KNN, SVM, SASOM, ensemble classifier

1 Introduction

The need to study whole genome such as Human Genomic Project (HGP) is recently increasing because fragmentary knowledge about life phenomenon with complex control functions of molecular-level is limited. DNA chips have been developed during that process because understanding the functions of genome sequences is essential at that time.

The development of DNA microarray technology has been produced large amount of gene data and has made it easy to monitor the expression patterns of thousands of genes simultaneously under particular experimental environments and conditions (Harrington *et al*. 2000). Also, we can analyze the gene information very rapidly and precisely by managing them at one time (Eisen *et al*. 1999).

Microarray technology has been applied to the field of accurate prediction and diagnosis of cancer and expected that it would help them. Especially accurate classification of cancer is very important issue for treatment of cancer. Many researchers have been studying many problems of cancer classification using gene expression profile data and attempting to propose the optimal classification technique to work out these problems (Dudoit *et al*. 2000, Ben-Dor *et al*. 2000) as shown in Table Ⅰ. Some produce better results than others, but there have been still no comprehensive work to compare the possible feature selection methods and classifiers. We need a thorough effort to give the evaluation of the possible methods to solve the problems of analyzing gene expression data.

The gene expression data usually consist of huge number of genes, and the necessity of tools analysing them to get useful information gets radical. There is research that systematically analyzes the results of test using a variety of feature selection methods and classifiers for selecting informative genes to help classification of cancer and classifying cancer (Ryu *et al*. 2002). However, the results were not verified enough because only one benchmark dataset was used. Due to the reason, it is necessary to analyse systematically the performance of classifiers using a variety of benchmark datasets.

In this paper, we attempt to explore many features and classifiers that precisely classify cancer using three recently published benchmark dataset. We adopted seven feature selection methods and four classifiers, which are commonly used in the field of data mining and pattern recognition. Feature selection methods include Pearson's and Spearman's correlation coefficients, Euclidean distance, cosine coefficient, information gain, mutual information and signal to noise ratio. Also, classification methods include multi-layer perceptron (MLP), k-nearest neighbour (KNN), support vector machine (SVM) and structure adaptive self –organizing map (SOM). We also attempt to combine some of the classifiers with majority voting to improve the performance of classification.

2 Backgrounds

2.1 DNA Microarray

DNA arrays consist of a large number of DNA molecules spotted in a systemic order on a solid substrate. Depending on the size of each DNA spot on the array, DNA arrays can be categorized as microarrays when the diameter of DNA spot is less than 250 microns, and macroarrays when the diameter is bigger than 300 microns. The arrays with the small solid substrate are also

referred to as DNA chips. It is so powerful that we can investigate the gene information in short time, because at least hundreds of genes can be put on the DNA microarray to be analyzed.

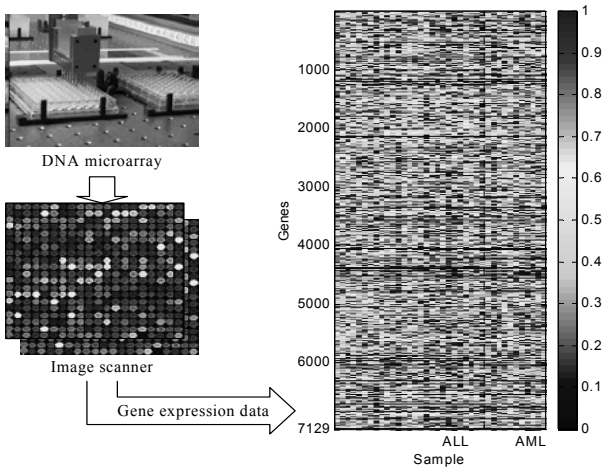

Fig. 1. General process of acquiring the gene expression data from DNA microarray

DNA microarrays are composed of thousands of individual DNA sequences printed in a high density array on a glass microscope slide using a robotic arrayer as shown in Fig. 1. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples are reverse-transcribed into cDNA, labeled using different fluorescent dyes mixed (red-fluorescent dye Cy5 and green-fluorescent dye Cy3). After the hybridization of these samples with the arrayed DNA probes, the slides are imaged using scanner that makes fluorescence measurements for each dye. The log ratio between the two intensities of each dye is used as the gene expression data (Lashkari *et al*. 1997, Derisi *et al*. 1997, Eisen *et al*. 1998).

$$gene\_expression = \log_2 \frac{\text{Int(Cy5)}}{\text{Int(Cy3)}} \qquad (1)$$

where Int(Cy5) and Int(Cy3) are the intensities of red and green colors. Since at least hundreds of genes are put on the DNA microarray, it is so helpful that we can

Table Ⅰ. Relevant works on cancer classification

Authors	Dataset	Method		Accuracy [%]
		Feature	Classifier	
Furey *et al.*	**Leukemia**	**Signal to noise ratio**	**SVM**	**94.1**
	Colon			**90.3**
Li *et al.* 2000	**Leukemia**	**Model selection with Akaike information criterion and Bayesian information criterion with logistic regression**		**94.1**
Li *et al.* 2001	**Lymphoma**	**Genetic Algorithm**	**KNN**	**84.6~**
	Colon			**94.1~**
Ben-Dor *et al.*	**Leukemia**	**All genes, TNoM score**	**Nearest neighbor**	**91.6**
	Colon			**80.6**
	Leukemia		**SVM with quadratic kernel**	**94.4**
	Colon			**74.2**
	Leukemia		**AdaBoost**	**95.8**
	Colon			**72.6**
Dudoit *et al.*	**Leukemia**	**The ratio of between-groups to within-groups sum of squares**	**Nearest neighbor**	**95.0~**
	Lymphoma			**95.0~**
	Leukemia		**Diagonal linear discriminant analysis**	**95.0~**
	Lymphoma			**95.0~**
	Leukemia		**BoostCART**	**95.0~**
	Lymphoma			**90.0~**
Nguyen *et al.*	**Leukemia**	**Principal component analysis**	**Logistic discriminant**	**94.2**
	Lymphoma			**98.1**
	Colon			**87.1**
	Leukemia		**Quadratic discriminant analysis**	**95.4**
	Lymphoma			**97.6**
	Colon			**87.1**
	Leukemia	**Partial least square**	**Logistic discriminant**	**95.9**
	Lymphoma			**96.9**
	Colon			**93.5**
	Leukemia		**Quadratic discriminant analysis**	**96.4**
	Lymphoma			**97.4**
	Colon			**91.9**

investigate the genome-wide information in short time.

2.2 Related Works

It is essential to efficiently analyze DNA microarray data because the amount of DNA microarray data is usually very large. The analysis of DNA microarray data is divided into four branches: clustering, classification, gene identification, and gene regulatory network modeling. Many machine learning and data mining methods have been applied to solve them.

Information theory (Fuhrman *et al*. 2000) has been applied to gene identification problem. Also, boolean network (Thieffry *et al*. 1998), Bayesian network (Friedman *et al*. 2000), and reverse engineering method (Arkin et al. 1997) have been applied to gene regulatory network modeling problem.

Several machine learning techniques have been previously used in classifying gene expression data, including Fisher linear discriminant analysis (Dudoit *et al*. 2000), *k* nearest neighbour (Li *et al*. 2001), decision tree, multi-layer perceptron (Khan *et al*. 2001, Xu *et al*. 2002), support vector machine (Furey et al. 2000, Brown *et al*. 2000), boosting, and self-organizing map (Golub *et al*. 1999). Also, many machine learning techniques were have been used in clustering gene expression data (Shamir 2001). They include hierarchical clustering (Eisen *et al*. 1998), self-organizing map (Tamayo *et al*. 1999), and graph theoretic approaches (Hartuv *et al*. 2000, Ben-Dor *et al*. 1999, Sharan et al. 2000)

The first approach, classification method, is called *supervised* method while the second approach, clustering method, is called *unsupervised* method. Clustering methods do not use any tissue annotation (e.g., tumor vs. normal) in the partitioning step. In contrast, classification methods attempt to predict the classification of new tissues, based on their gene expression profiles after training on examples (training data) that have been classified by an external "supervision" (Ben-Dor *et al*. 2000). Table Ⅰ shows relevant works on cancer classification.

3 Machine Learning for DNA Microarray

We define machine learning for DNA microarray that selects discriminative genes related with classification from gene expression data, trains classifier and then classifies new data using learned classifier. The system is as shown in Fig. 2. After acquiring the gene expression data calculated from the DNA microarray, our prediction system has 2 stages: feature selection and pattern classification stages.

The feature selection can be thought of as the gene selection, which is to get the list of genes that might be informative for the prediction by statistical, information theoretical methods, etc. Since it is highly unlikely that all the 7,129 genes have the information related to the cancer and using all the genes results in too big dimensionality, it is necessary to explore the efficient way to get the best feature. We have extracted 25 genes using seven methods described in Section 3.1, and the cancer predictor classifies the category only with these genes.

Given the gene list, a classifier makes decision to which category the gene pattern belongs at prediction stage. We have adopted four most widely used classification methods and an ensemble classifier as shown in Fig. 2.

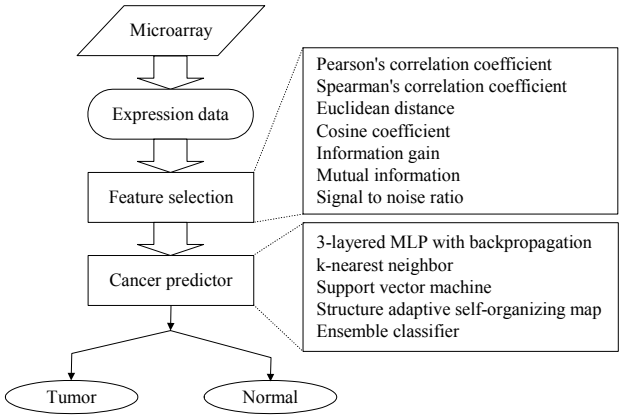

Fig. 2. Cancer classification system

3.1 Gene Selection

Among thousands of genes whose expression levels are measured, not all are needed for classification. Microarray data consist of large number of genes in small samples. We need to select some genes highly related with particular classes for classification, which is called informative genes (Golub *et al*. 1999). This process is referred to as gene selection. It is also called feature selection in machine learning.

Using the statistical correlation analysis, we can see the linear relationship and the direction of relation between two variables. Correlation coefficient r varies from –1 to +1, so that the data distributed near the line biased to (+) direction will have positive coefficients, and the data near the line biased to (-) direction will have negative coefficients.

Suppose that we have a gene expression pattern $\boldsymbol{g_i}$ (i = 1 ~ 7,129 in Leukemia data, i = 1 ~ 2,000 in Colon data, i = 1 ~ 4,026 in Lymphoma data). Each $\boldsymbol{g_i}$ is a vector of gene expression levels from N samples, $\boldsymbol{g_i} = (e_1, e_2, e_3, \ldots, e_N)$. The first M elements ($e_1, e_2, \ldots, e_M$) are examples of tumor samples, and the other N-M ($e_{M+1}, e_{M+2}, \ldots, e_N$) are those from normal samples. An ideal gene pattern that belongs to tumor class is defined by $g_{ideal\_tumor} = (1, 1, \ldots, 1, 0, \ldots, 0)$, so that all the elements from tumor samples are 1 and the others are 0. In this paper, we have calculated the correlation coefficient between this g_{ideal} and the expression pattern of each gene. When we have two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ that contain N elements, $r_{Pearson}$ and $r_{Spearman}$ are calculated as follows:

$$r_{Pearson} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} \quad (2)$$

$$r_{Spearman} = 1 - \frac{6\sum (D_x - D_y)^2}{N(N^2 - 1)} \quad (3)$$

where, D_x and D_y are the rank matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

The similarity between two input vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be thought of as distance. Distance is a measure on how far the two vectors are located, and the distance between g_{ideal_tumor} and g_i tells us how much the g_i is likely to the tumor class. Calculating the distance between them, if it is bigger than certain threshold, the gene g_i would belong to tumor class, otherwise g_i belongs to normal class. In this paper, we have adopted Euclidean distance ($r_{Eclidean}$) and cosine coefficient (r_{Cosine}) represented by the following equations:

$$r_{Eclidean} = \sqrt{\sum (X - Y)^2} \quad (4)$$

$$r_{Cosine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad (5)$$

We have utilized the information gain and mutual information that are widely used in many fields such as text categorization and data mining. If we count the number of genes excited ($P(g_i)$) or not excited ($P(\bar{g}_i)$) in category c_j ($P(c_j)$), the coefficients of the information gain and mutual information become as follows:

$$IG(g_i, c_j) = P(g_i, c_j) \log \frac{P(g_i, c_j)}{P(c_j) \cdot P(g_i)} + P(\bar{g}_i, c_i) \log \frac{P(\bar{g}_i, c_j)}{P(c_j) \cdot P(\bar{g}_i)} \quad (6)$$

$$MI(g_i, c_j) = \log \frac{P(g_i, c_j)}{P(c_j) \cdot P(g_i)} \quad (7)$$

Mutual information tells us the dependency relationship between two probabilistic variables of events. If two events are completely independent, the mutual information is 0. The more they are related, the higher the mutual information gets. Information gain is used when the features of samples are extracted by inducing the relationship between gene and class by the presence frequency of the gene in the sample. Information gain measures the goodness of gene using the presence and absence within the corresponding class.

For each gene g_i, some are from tumor samples, and some are from normal samples. If we calculate the mean μ and standard deviation σ from the distribution of gene expressions within their classes, the signal to noise ratio of gene g_i, $SN(g_i)$, is defined by:

$$SN(g_i) = \frac{\mu_{tumor}(g_i) - \mu_{normal}(g_i)}{\sigma_{tumor}(g_i) - \sigma_{normal}(g_i)} \quad (8)$$

3.2 Classification

Many algorithms designed for solving classification problems in machine learning have been applied to recent research of prediction and classification of cancer with gene expression data. General process of classification in machine learning is to train classifier to accurately recognize patterns from given training samples and to classify test samples with the trained classifier. Representative classification algorithms such as multi-layer perceptron, *k*-nearest neighbour, support vector machine, and structure-adaptive self-organizing map are applied to the classification.

1) MLP

Error backpropagation neural network is a feed-forward multilayer perceptron (MLP) that is applied in many fields due to its powerful and stable learning algorithm (Lippman *et al*. 1987). The neural network learns the training examples by adjusting the synaptic weight of neurons according to the error occurred on the output layer. The power of the backpropagation algorithm lies in two main aspects: local for updating the synaptic weights and biases, and efficient for computing all the partial derivatives of the cost function with respect to these free parameters (Beale 1996). The weight-update rule in backpropagation algorithm is defined as follows:

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1) \quad (9)$$

where $\Delta w_{ji}(n)$ is the weight update performed during the *n*th iteration through the main loop of the algorithm, η is a positive constant called the learning rate, δ_j is the error term associated with j, x_{ji} is the input from node i to unit j, and $0 \leq \alpha < 1$ is a constant called the *momentum*.

2) KNN

k-nearest neighbor (KNN) is one of the most common methods among memory based induction. Given an input vector, KNN extracts *k* closest vectors in the reference set based on similarity measures, and makes decision for the label of input vector using the labels of the *k* nearest neighbors.

Pearson's coefficient correlation and Euclidean distance have been used as the similarity measure. When we have an input X and a reference set $D = \{d_1, d_2, \cdots, d_N\}$, the probability that X may belong to class c_j, $P(X, c_j)$ is defined as follows:

$$P(X, c_j) = \sum_{d_i \in kNN} \mathrm{Sim}(X, d_i) P(d_i, c_j) - b_j \quad (10)$$

where $\mathrm{Sim}(X, d_i)$ is the similarity between X and d_i and b_j is a bias term.

3) SASOM

Self-organizing map (SOM) defines a mapping from the input space onto an output layer by unsupervised learning algorithm. SOM has an output layer consisting of N nodes, each of which represents a vector that has the same dimension as the input pattern. For a given input vector X, the winner node m_c is chosen using Euclidean distance between $\boldsymbol{x}$ and its neighbors, $\boldsymbol{m_i}$.

$$\|x - m_c\| = \min_i \|x - m_i\| \tag{11}$$

$$m_i(t+1) = m_i(t) + \alpha(t) \times n_{ci}(t) \times \{x(t) - m_i(t)\} \tag{12}$$

Even though SOM is well known for its good performance of topology preserving, it is difficult to apply it to practical classification since the topology should be fixed before training. A structure adaptive self-organizing map (SASOM) is proposed to overcome this shortcoming (Kim *et al*. 2000). SASOM starts with 4×4 map, and dynamically splits the output nodes of the map, where the data from different classes are mixed, trained with the LVQ learning algorithm. Fig. 3 illustrates the algorithm of SASOM.

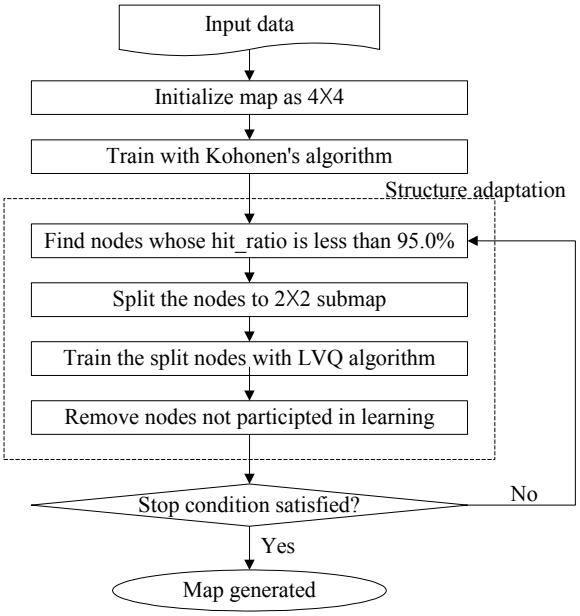

Fig. 3. Overview of SASOM

4) SVM

Support vector machine (SVM) estimates the function classifying the data into two classes (Vapnik 1995, Moghaddam *et al*. 2000). SVM builds up a hyperplane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. SVM achieves this by the structural risk minimization principle that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension. Given a labeled set of M training samples ($\boldsymbol{X_i}$, $\boldsymbol{Y_i}$), where $\boldsymbol{X_i} \in R^N$ and $\boldsymbol{Y_i}$ is the associated label, $Y_i \in \{-1, 1\}$, the discriminant hyperplane is defined by:

$$f(X) = \sum_{i=1}^{M} Y_i \alpha_i k(X, X_i) + b \tag{13}$$

where $k(\ .)$ is a kernel function and the sign of $f(X)$ determines the membership of X. Constructing an optimal hyperplane is equivalent to finding all the nonzero α_i (support vectors) and a bias b. We have used SVM$^{\text{light}}$ module and SVM$^{\text{RBF}}$ in this paper.

5) Ensemble classifier

Classification can be defined as the process to approximate I/O mapping from the given observation to the optimal solution. Generally, classification tasks consist of two parts: feature selection and classification. Feature selection is a transformation process of observations to obtain the best pathway to get to the optimal solution. Therefore, considering multiple features encourages obtaining various candidate solutions, so that we can estimate more accurate solution to the optimal than any other local optima.

When we have multiple features available, it is important to know which of features should be used. Theoretically, as many features we may concern, it may be more effective for the classifier to solve the problems. But features that have overlapped feature spaces may cause the redundancy of irrelevant information and result in the counter effect such as overfitting. Therefore, it is more important to explore and utilize independent features to train classifiers, rather than increase the number of features we use. Correlation between feature sets can be induced from the distribution of feature numbers, or using mathematical analysis using statistics.

Meanwhile, there are many algorithms for the classification from machine learning approach, but none of them is perfect. However, it is always difficult to decide what to use and how to set up its parameters. According to the environments the classifier is embedded, some algorithm works well and others not. It is because, depending on the algorithms, features and parameters used, the classifier searches in different solution space. These sets of classifiers produce their own outputs, and enable the ensemble classifier to explore more wide solution space.

We have applied this idea to a classification framework as shown in Fig. 4. If there are k features and n classifiers, there are $k \times n$ feature-classifier combinations. There are ${}_{k\times n}C_m$ possible ensemble classifiers when m feature-classifier combinations are selected for ensemble classifier. Then classifiers are trained using the features selected, finally a majority voting is accompanied to combine the outputs of these classifiers. After classifiers with some features are trained independently produce their own outputs, final answer will be judged by a combining module, where the majority voting method is adopted.

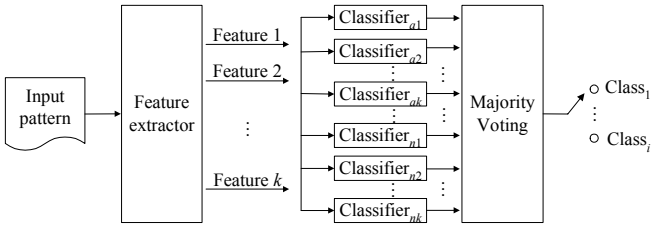

Fig 4. Overview of the ensemble classifier

4 Experimental Results

4.1 Datasets

There are several microarray datasets from published cancer gene expression studies, including leukemia cancer dataset, colon cancer dataset, lymphoma dataset, breast cancer dataset, NCI60 dataset, and ovarian cancer dataset. Among them three datasets are used in this paper. The first dataset and third dataset involve samples from two variants of the same disease and second dataset involves tumor and normal samples of the same tissue. Because the benchmark data have been studied in many papers, we can compare the results of this paper with others.

1) Leukemia cancer dataset

Leukemia dataset consists of 72 samples: 25 samples of *acute myeloid leukemia* (AML) and 47 samples of *acute lymphoblastic leukemia* (ALL). The source of the gene expression measurements was taken form 63 bone marrow samples and 9 peripheral blood samples. Gene expression levels in these 72 samples were measured using high density oligonucleotide microarrays (Ben-Dor *et al*. 2000).

38 out of 72 samples were used as training data and the remaining were used as test data in this paper. Each sample contains 7129 gene expression levels.

2) Colon cancer dataset

Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. Although original data consists of 6000 gene expression levels, 4000 out of 6000 were removed based on the confidence in the measured expression levels. 40 of 62 samples are colon cancer samples and the remaining are normal samples. Each sample was taken from tumors and normal healthy parts of the colons of the same patients and measured using high density oligonucleotide arrays (Ben-Dor *et al*. 2000).

31 out of 62 samples were used as training data and the remaining were used as test data in this paper.

3) Lymphoma cancer dataset

B cell diffuse large cell lymphoma (B-DLCL) is a heterogeneous group of tumors, based on significant variations in morphology, clinical presentation, and response to treatment. Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal center B cell-like DLCL and activated B cell-like DLCL (Lossos *et al*. 2000). Lymphoma dataset consists of 24 samples of GC B-like and 23 samples of activated B-like.

22 out of 47 samples were used as training data and the remaining were used as test data in this paper.

4.2 Environments

For feature selection, each gene is scored based on the feature selection methods described in Section 3.1, and the 25 top-ranked genes are chosen as the feature of the input pattern.

For classification, we have used 3-layered MLP with 5~15 hidden nodes, 2 output nodes, 0.01~0.50 of learning rate and 0.9 of momentum. KNN has been used with k=1~8. Similarity measures used in KNN are Pearson's correlation coefficient and Euclidean distance. SASOM has been used by 4×4 map with rectangular topology, 0.05 of initial learning rate, 1000 of initial learning length, 10 of initial radius, 0.02 of final learning rate, 10000 of final learning length and 3 of final radius. We have used SVM with linear function and RBF function as kernel function. In RBF, we have changed 0.1~0.5 gamma variable.

4.3 Analysis of results

Table Ⅱ shows the IDs of genes overlapped by Pearson's correlation coefficient, cosine coefficient, Euclidean distance in each dataset. Among these genes there are some genes overlapped by other feature selection methods. For example, gene 2288 of leukemia has been third-ranked in information gain. The number of overlapped genes of leukemia dataset is 17. The number of overlapped genes of colon dataset is 9. The number of overlapped genes of lymphoma dataset is 19. These overlapped genes are very informative. In particular, Zyxin, gene 4847 of leukemia, has been reported as informative (Golub *et al*. 1999), but there are no genes appeared commonly in every method.

Table Ⅱ. The IDs of genes overlapped by Pearson's correlation coefficient, cosine coefficient, and Euclidean distance

Leukemia	**461**	**1249**	**1745**	**1834**	**2020**
	2043	**2242**	**2288**	**3258**	**3320**
	4196	**4847**	**5039**	**6200**	**6201**
	6373	**6803**			
Colon	**187**	**619**	**704**	**767**	**1060**
	1208	**1546**	**1771**	**1772**	
Lymphoma	**36**	**75**	**76**	**77**	**86**
	86	**678**	**680**	**1636**	**1637**
	2225	**2243**	**2263**	**2412**	**2417**
	2467	**3890**	**3893**	**3934**	

Fig. 5 shows the expression level of genes chosen by Pearson's correlation coefficient method in Leukemia dataset. 1~27 samples are ALL and 28~38 samples are AML. The differences of brightness between AML and ALL represent that genes chosen by Pearson's correlation coefficient method divide samples into AML and ALL.

The results of recognition rate on the test data are as shown in Tables Ⅲ, Ⅳ, and Ⅴ. Column is the list of feature selection methods: Pearson's correlation

coefficient (PC), Spearman's correlation coefficient (SC), Euclidean distance (ED), cosine coefficient (CC), information gain (IG), mutual information (MI), and signal to noise ratio (SN). $KNN_{Pearson}$ and MLP seem to produce the best recognition rate among the classifiers on the average. $KNN_{Pearson}$ is better than KNN_{cosine}. SVM is poorer than any other classifiers.

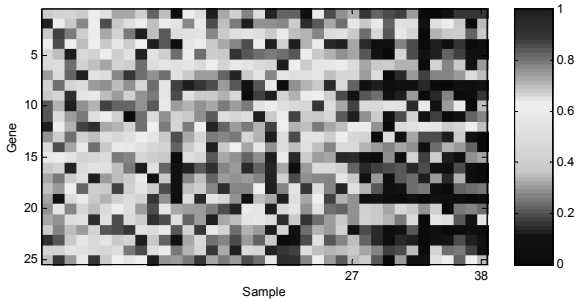

Fig. 5. Expression level of genes chosen by $r_{Pearson}$ in Leukemia dataset

Table Ⅲ. Recognition rate with features and classifiers (%) in Leukemia dataset

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	97.1	76.5	79.4	79.4	97.1	94.1
SC	82.4	61.8	58.8	58.8	76.5	82.4
ED	91.2	73.5	70.6	70.6	85.3	82.4
CC	94.1	88.2	85.3	85.3	91.2	94.1
IG	97.1	91.2	97.1	97.1	94.1	97.1
MI	58.8	58.8	58.8	58.8	73.5	73.5
SN	76.5	67.7	58.8	58.8	73.5	73.5
Mean	85.3	74.0	72.7	72.7	84.5	85.3

Table Ⅳ. Recognition rate with features and classifiers (%) in Colon dataset

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	74.2	74.2	64.5	64.5	71.0	77.4
SC	58.1	45.2	64.5	64.5	61.3	67.7
ED	67.8	67.6	64.5	64.5	83.9	83.9
CC	83.9	64.5	64.5	64.5	80.7	80.7
IG	71.0	71.0	71.0	71.0	74.2	80.7
MI	71.0	71.0	71.0	71.0	74.2	80.7
SN	64.5	45.2	64.5	64.5	64.5	71.0
Mean	70.1	62.7	66.4	66.4	72.7	77.4

Table Ⅴ. Recognition rate with features and classifiers (%) in Lymphoma dataset

	MLP	SASOM	SVM		KNN	
			Linear	RBF	Cosine	Pearson
PC	64.0	48.0	56.0	60.0	60.0	76.0
SC	60.0	68.0	44.0	44.0	60.0	60.0
ED	56.0	52.0	56.0	56.0	56.0	68.0
CC	68.0	52.0	56.0	56.0	60.0	72.0
IG	92.0	84.0	92.0	92.0	92.0	92.0
MI	72.0	64.0	64.0	64.0	80.0	64.0
SN	76.0	76.0	72.0	76.0	76.0	80.0
Mean	69.7	63.4	62.9	63.4	69.1	73.1

Fig. 6 shows the comparison of the average performance of features. Although the results are different between datasets, information gain is the best, and Pearson's correlation coefficient is the second. Mutual information and Spearman's correlation coefficient are poor. The difference of performance in datasets might be caused by the characteristics of data.

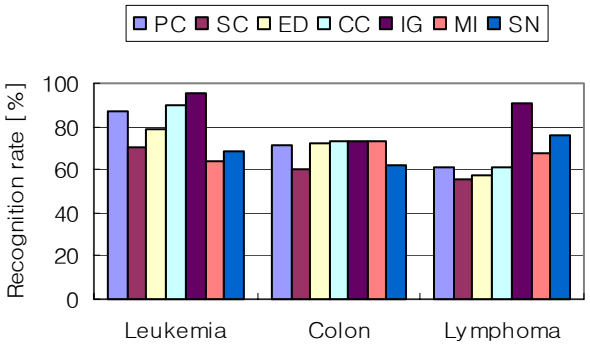

Fig. 6. Average performance of feature selection methods

Recognition rates by ensemble classifiers are shown in Table Ⅵ. Majority voting-3 means the ensemble classifier using majority voting with 3 classifiers, and majority voting-all means the ensemble classifier using majority voting with all 42 feature-classifier combinations. Fig. 7 shows the comparison of the performance of the best classifier of all possible ${}_{42}C_3$ ensemble classifiers, ensemble classifier-3 and ensemble classifier-all. The best result of Leukemia is obtained by all classifier except SASOM. The result of the best classifier is the same as that of the best ensemble classifier using majority voting with 3 classifiers. In other datasets, the performance of ensemble classifier surpasses the best classifier. In all datasets, ensemble classifier using majority voting with all classifiers are the worst.

Table Ⅵ. Recognition rate by ensemble classifier

	Majority voting-3	Majority voting-all
Leukemia	97.1	91.2
Colon	93.6	71.0
Lymphoma	96.0	80.0

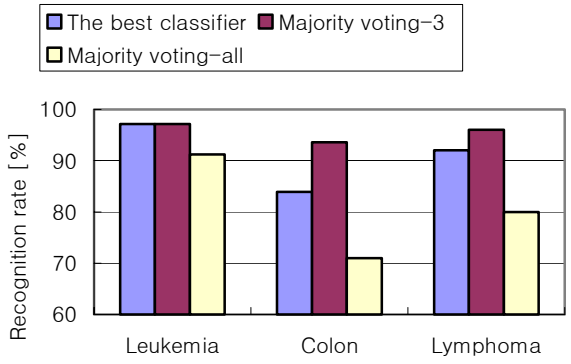

Fig. 7. Comparison of the performance of the best classifier, the best ensemble classifier-3, and ensemble classifier-all

Table Ⅶ. Classifiers of the best ensemble classifier of all possible ${}_{42}C_3$ ensemble classifiers in Colon dataset

Classifier	Feature selection method
MLP	**Cosine coefficient**
KNN_{cosine}	**Euclidean distance**
KNN_{cosine}	**Pearson's correlation coefficient**
MLP	**Cosine coefficient**
KNN_{cosine}	**Euclidean distance**
$KNN_{Pearson}$	**Pearson's correlation coefficient**
MLP	**Cosine coefficient**
KNN_{cosine}	**Euclidean distance**
SASOM	**Pearson's correlation coefficient**
MLP	**Mutual information**
KNN_{cosine}	**Euclidean distance**
$KNN_{pearson}$	**Pearson's correlation coefficient**
MLP	**Information gain**
KNN_{cosine}	**Euclidean distance**
$KNN_{pearson}$	**Pearson's correlation coefficient**
MLP	**Cosine coefficient**
MLP	**Pearson's correlation coefficient**
$KNN_{pearson}$	**Euclidean distance**
$KNN_{pearson}$	**Euclidean distance**
$KNN_{pearson}$	**Mutual information**
SASOM	**Pearson's correlation coefficient**
$KNN_{pearson}$	**Euclidean distance**
$KNN_{pearson}$	**Information gain**
SASOM	**Pearson's correlation coefficient**

Table Ⅶ shows the classifiers of the best ensemble classifier of all possible ${}_{42}C_3$ ensemble classifiers in Colon dataset where its recognition rate is 93.6%. If we observe the classifiers of the best ensemble classifier in Fig. 10, we find features more important to affect the result than classifiers. In other words, in ensemble classifiers there must be classifiers with Euclidean distance and Pearson's correlation coefficient. The other classifier is the one with cosine coefficient, mutual information or information gain.

This fact is also prominent in Lymphoma dataset. Most of the classifiers of the best ensemble classifiers are classifiers with information gain, signal to noise ratio and Euclidean distance, or the classifiers with information gain, signal to noise ratio and Pearson's correlation coefficient.

As shown in Fig. 8~11, Euclidean distance, Pearson's correlation coefficient and cosine coefficient are highly correlated in Colon dataset. Table Ⅷ shows genes ranked by Euclidean distance, Pearson's correlation coefficient and cosine coefficient and the value of genes by each method. The bold faced figures mean the overlapped genes of those features. There are some overlapped genes among them as shown in Table Ⅱ. This indicates overlapped genes of highly correlated features can discriminate classes and the other genes not overlapped among combined features can supplement to search the solution spaces. For example, gene 1659 and gene 550 are high-ranked in both of Pearson's correlation coefficient and cosine coefficient, and gene 440 is high-ranked in both of Euclidean distance and cosine coefficient. This subset of two features might paly an important role in classification.

This paper shows that the ensemble classifier works and we can improve the classification performance by combining complementary common sets of classifiers learned from three independent features, even when we use simple combination method like majority voting.

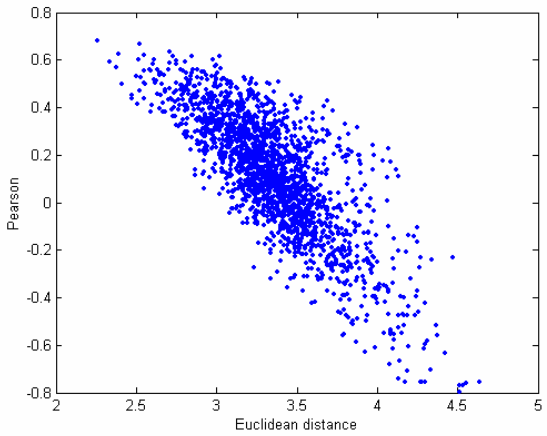

Fig. 8. Correlation of Euclidean distance and Pearson's correlation coefficient in Colon dataset

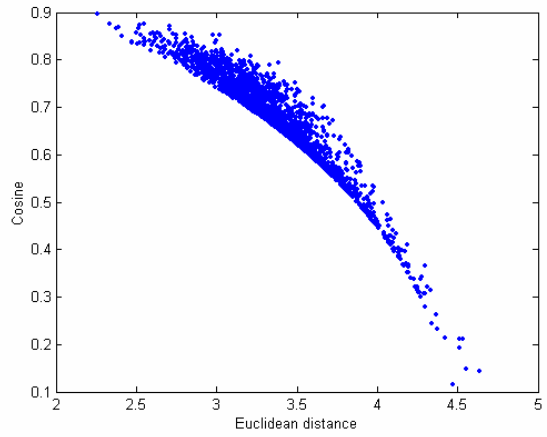

Fig. 9. Correlation of Euclidean distance and cosine coefficient in Colon dataset

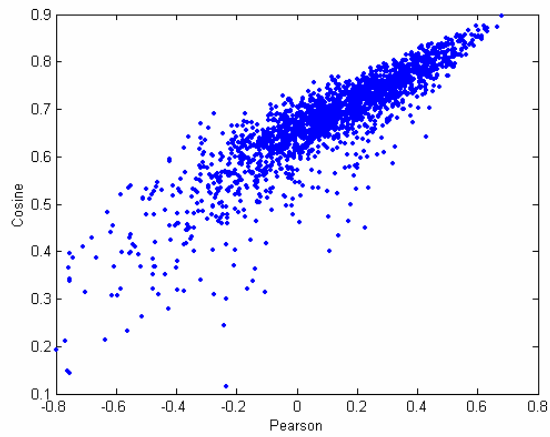

Fig. 10. Correlation of Pearson's correlation coefficient and cosine coefficient in Colon dataset

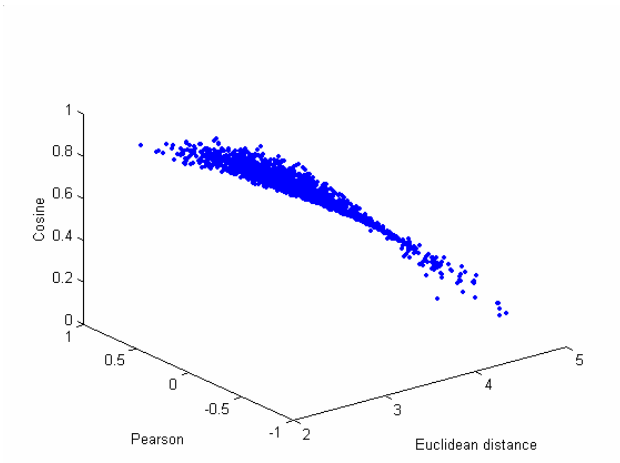

Fig. 11. Correlation of Euclidean distance, Pearson's correlation coefficient and cosine coefficient in Colon dataset

Table Ⅷ. Genes ranked by Euclidean distance, Pearson's correlation coefficient and cosine coefficient

Rank	Euclidean	Pearson	Cosine
1	**619**(2.262385)	**619**(0.681038)	**619**(0.895971)
2	**767**(2.335303)	**1771**(0.664378)	**1772**(0.875472)
3	**704**(2.374358)	1659(0.634084)	**767**(0.874914)
4	**187**(2.388404)	550(0.631655)	**1771**(0.873892)
5	207(2.410640)	**187**(0.626262)	1659(0.870115)
6	887(2.473033)	**1772**(0.621581)	**187**(0.867285)
7	635(2.474971)	1730(0.615566)	**704**(0.866679)
8	1915(2.498611)	1648(0.614949)	**1208**(0.866029)
9	1046(2.506833)	365(0.614591)	550(0.864547)
10	**1208**(2.512257)	**1208**(0.603313)	**1546**(0.856904)
11	482(2.520699)	1042(0.602160)	251(0.855841)
12	**1771**(2.525080)	**1060**(0.601712)	1915(0.855784)
13	1993(2.529032)	513(0.596444)	440(0.855453)
14	62(2.546894)	**767**(0.594119)	1263(0.854854)
15	**1772**(2.547455)	1263(0.591725)	**1060**(0.854829)
16	1194(2.549244)	138(0.587851)	965(0.854137)
17	1594(2.551892)	1826(0.584774)	1648(0.854119)
18	199(2.557360)	**1546**(0.582293)	1942(0.853586)
19	1867(2.587469)	141(0.579073)	513(0.852270)
20	959(2.589989)	1227(0.574537)	1042(0.851993)
21	440(2.593881)	**704**(0.569022)	1993(0.851753)
22	480(2.594514)	1549(0.562828)	365(0.851205)
23	**1546**(2.604907)	1489(0.561003)	1400(0.849531)
24	399(2.613609)	1724(0.559919)	207(0.849084)
25	**1060**(2.614100)	1209(0.559778)	271(0.848481)

5 Concluding Remarks

We have conducted a thorough quantitative comparison among the 42 combinations of features and classifiers for three benchmark dataset. Information gain and Pearson's correlation coefficient are the top feature selection methods, and MLP and KNN are the best classifiers. The experimental results also imply some correlations between features and classifiers, which might guide the researchers to choose or devise the best classification method for their problems in bioinformatics. Based on the results, we have developed the optimal feature-classifier combination to produce the best performance on the classification.

We have combined 3 classifiers among 42 classifiers using majority voting. We could confirm that ensemble classifier of highly correlated features works better than ensemble of uncorrelated features. In particular, we analyzed the improvement of the classification performance for Colon dataset.

Moreover, our method of combining classifiers is very simple, and there are many methods of combining classifiers in machine learning and data mining fields. We will have to apply more sophisticated methods of combining classifiers to the same dataset to confirm the results obtained and get better results.

Acknowledgements

This paper was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Science and Technology.

6 References

Alon, U., Barkai, N., et al. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. of the Natl. Acad. of Sci. USA*, **96**:6745-6750.

Arkin, A., Shen, P. and Ross, J. (1997): A test case of correlation metric construction of a reaction pathway from measurements. *Science*, **277**:1275-1279.

Beale, H. D. (1996): *Neural Network Design*. 1-47, PWS Publish Company.

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999): Clustering gene expression patterns. *Journal of Computational Biology*, **6**:281-297.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, N. (2000): Tissue classification with gene expression profiles. *Journal of Computational Biology*, **7**:559-584.

Brown, M. P. S., 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. of the Natl. Acad. of Sci. USA*, **97**:262-267, 2000.

Derisi, J., Iyer, V. and Brosn, P. (1997): Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**:680-686.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2000): Comparison of discrimination methods for the classification of tumors using gene expression data. *Technical Report 576*, Department of Statistics, University of California, Berkeley.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Bostein, D. (1998): Cluster analysis and display of genome-wide expression patterns. *Proc. of the Natl. Acad. of Sci. USA*, **95**:14863-14868.

Eisen, M. B. and Brown, P. O. (1999): DNA arrays for analysis of gene expression. *Methods Enzymbol*, **303**: 179-205.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000): Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**:601-620.

Fuhrman, S., Cunningham, M. J., Wen, X., Zweiger, G., Seilhamer, J. and Somogyi, R. (2000): The application of Shannon entropy in the identification of putative drug targets. *Biosystems*, **55**:5-14.

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**(10):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., Blomfield, C. D., and Lander, E. S. (1999): Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring. *Science*, **286**:531-537.

Harrington, C. A., Rosenow, C., and Retief, J. (2000): Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.*, **3**:285-291.

Hartuv, E., Schmitt, A., Lange, J., Meier-Ewert, S., Lehrach, H. and Shamir, R. (2000): An algorithm for clustering cDNA fingerprints. *Genomics*, **66**(3):249-256.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. And Meltzer, P. S. (2001): Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**(6):673-679.

Kim, H. D. and Cho, S.-B. (2000): Genetic optimization of structure-adaptive self-organizing map for efficient classification. *Proc. of International Conference on Soft Computing*, 34-39, World-Scientific Publishing.

Lashkari, D., Derisi, J., McCusker, J., Namath, A., Gentile, C., Hwang, S., Brown, P., and Davis, R. (1997): Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. of the Natl. Acad. of Sci. USA*, **94**:13057-13062.

Lippman, R. P. (1987): An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4-22.

Li, L., Weinberg, C. R., Darden, T. A. and Pedersen, L. G. (2001): Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**(12):1131-1142.

Li, W. and Yang, Y. (2000): How many genes are needed for a discriminant microarray data analysis. *Critical Assessment of Techniques for Microarray Data Mining Workshop*.

Lossos, I. S., Alizadeh, A. A., Eisen, M. B., Chan, W. C., Brown, P. O., Bostein, D., Staudt, L. M., and Levy, R. (2000): Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas. *Proc. of the Natl. Acad. of Sci. USA*, **97**(18):10209-10213.

Moghaddam, B. and Yang, M.-H. (2000): Gender classification with support vector machines. *Proc. of 4th IEEE Intl. Conf. on Automatic Face and Gesture Recognition*, 306-311.

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

Quinlan, J. R. (1986): The effect of noise on concept learning. *Machine Learning: An Artificial Intelligence Approach*, Michalski, R. S., Carbonell, J. G. and Mitchell, T. M. (eds). San Mateo, CA: Morgan Kauffmann, **2**:149-166.

Ryu, J. and Cho, S. B. (2002): Towards optimal feature and classifier for gene expression classification of cancer. *Lecture Note in Artificial Intelligence*, **2275**:310-317.

Shamir, R. and Sharan, R. (2001): Algorithmic approaches to clustering gene expression data. *Current Topics in Computational Biology*. In Jiang, T., Smith, T., Xu, Y. and Zhang, M. Q. (eds), MIT press.

Sharan, R. and Shamir, R. (2000): CLICK: A clustering with applications to gene expression analysis. *Proc. Of the Eighth International Conference in Computational Molecular Biology* (ISBM), 307-316.

Tamayo, P. (1999): Interpreting patterns of gene expression with self-organizing map: Methods and application to hematopoietic differentiation. *Proc. of the National Academy of Sciences of the United States of America*, **96**: 2907-2912.

Thieffry, D. and Thomas, R. (1998): Qualitative analysis of gene networks. *Pacific Symposium on Biocomputing*, **3**:66-76.

Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory*, New York: Springer.

Xu, Y., Selaru, M., Yin, J., Zou, T. T., Shustova, V., Mori, Y., Sato, F., Liu, T. C., Olaru, A., Wang, S., Kimos, M. C., Perry, K., Desai, K., Greenwood, B. D., Krasna, M. J., Shibata, D., Abraham, J. M. and Meltzer, S. J. (2002): Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Research*, **62**:3493-3497.