

# Finding expressed genes using genetic algorithms and support vector machines

Xue-wen Chen and Michael McKee

Department of Electrical and Computer Engineering, California State University  
18111 Nordhoff Street, Northridge, CA 91330, USA

Contact: [xwchen@csun.edu](mailto:xwchen@csun.edu)

## ABSTRACT

The gene expression data obtained from microarrays have shown useful in cancer classification. DNA microarray data have extremely high dimensionality compared to the small number of available samples. An important step in microarray studies is to remove genes irrelevant to the learning problem and to select a small number of genes expressed in biological samples under specific conditions. We propose a novel feature subset selection algorithm to identify expressed genes for cancer classification. This algorithm is based on genetic algorithms and support vector machine algorithms. This new algorithm is very efficient for selecting sets of genes in very high dimensional feature space. Two databases are considered: the colon cancer database and the leukemia database. Our experimental results show the effectiveness of the proposed algorithms.

**Keywords:** Feature selection, genetic algorithm, microarray data, support vector machines

## 1. INTRODUCTION

The advent of DNA microarray technology has made it possible to analyze thousands or tens of thousands of genes simultaneously (De Risi et al. 1997; Cho et al. 1998; Chu et al. 1998). This hybridization based technology revolutionizes the traditional ways in molecular biology (which works on one gene in one experiment) and is having a significant impact on genomics study. Information gained from microarray data allows for identifying genes that regulate key pathways in a cell, for extracting biological significance such as the changes in expression patterns of genes under different body tissues and different developmental stages, and for exploiting important clues to understanding the role of genes and the underlying gene regulatory network (Eisen et al. 1998; Eisen and Brown, 1999). Microarray technologies have found applications in many different areas such as gene discovery, disease diagnosis, and drug discovery.

Given such huge amount of data (tens or hundreds of data points for thousands or tens of thousands of genes), an important part in microarray studies is to make sense of the data and to draw biologically meaningful conclusions. This requires to remove genes irrelevant to the learning problems at hand and to select a small number of genes expressed in biological samples under specific conditions. The problem is known in the machine learning community as the feature subset selection problem (where each gene is considered as a feature). Feature selection is essential to reduce the test errors in microarray data processing as the number of genes is much larger than the number of available samples. Selecting expressed genes is particularly important in genomics studies. For example, in cancer classification and diagnosis, knowing when certain genes are expressed, which genes are suitable as marker genes, which genes are responsible for the change from normal to cancerous cells etc. can help understand the underlying molecular mechanisms and identify therapies targeted to different varieties of cancers.

A number of approaches have been recently proposed to select a subset of genes to use in microarray data analysis. Golub et al. (1999) developed a neighborhood analysis method to test the correlation between 6817 genes and to identify a subset of genes for cancer classification. Xing et al. (2001) used a statistical test called information gain to rank genes. Other statistical tests, including t-test with a Gaussian model (Long et al., 2001), the Fisher score (Furey et al., 2000), a hierarchical Gamma-Gamma-Bernoulli model (Newton et al., 2001), ANOVA A F-statistics (Kerr et al., 2000), and a nonparametric t-test (Dudoit et al., 2002), have also been applied to order individual genes. Essentially, all these methods consider each feature individually and rank features based on the power of each individual feature to separate samples with different class labels.

Due to the facts that genes do not work independently (an activated gene usually affects the expression levels of other genes) and several genes are turned on simultaneously at any given time, it is thus desirable exploring the information on interactions between genes. When one wishes to extract a subset of genes from a much larger gene set, the number of trials needed to evaluate this subset can be seen to be a combinatorial problem. To find the best subset of expressed genes, an exhaustive search is needed. However, exhaustive search is impractical or impossible for gene selection for microarray data because of the very high dimensional feature space (thousands or tens of thousands of genes). Other suboptimal search algorithms are thus employed. Bo and Jonassen (2002), and Inza et al. (2002) used a greedy search algorithm called sequential forward selection (SFS) to find expressed genes for classification. The SFS method first selects the best single feature and then adds one feature at a time which in combination with the selected features maximizes a criterion function. The SFS method is computationally attractive.

Most recently, Li et al. (2001) applied a stochastic search method, the genetic algorithm (GA), and a k-nearest neighbor (KNN) method to assess 50 genes for cancer classification. K-nearest neighbor decision rule was used to evaluate the fitness of gene expressions and a genetic algorithm was applied to identify 50 best genes that correctly classify all training samples. Due to their parallel processing capability and their ability to exploit the information adaptively, GAs have been shown to produce better performances than sequential selection methods (Raymer et al., 2000). GAs have been successfully applied to many different search and optimization problems.

In this paper, we present a novel approach for the selection of expressed genes using GAs and support vector machines (SVM). This paper differs from the GA/kNN approach (Li et al., 2001) mainly in two aspects. Firstly, the GA based gene selection is performed in the context of SVMs, instead of kNN. SVMs are attractive due to their ability to generalize well (Vapnik, 1998). This is particularly important in microarray data classifications, since only very limited training samples are available. Secondly, we present a different strategy using GAs. Both crossover and mutation will be operated with some modifications in GAs. The proposed GA/SVM algorithm is very efficient for selecting sets of genes in very high dimensional feature spaces for classification problems. Two databases are considered: the colon cancer database and the leukacmia database. Our experimental results show the effectiveness of the proposed GA/SVM algorithm.

The paper is organized as follows. In section 2, we describe our GA/SVM system. In section 3, we present the gene selection results on two microarray databases. We then draw conclusions in section 4.

## 2. GA/SVM METHODS

In this section, we describe the novel GA/SVM based gene selection methods. Our goal is to select a subset of predictive genes whose expressions can distinguish samples from different classes (e.g., tumor cells versus normal cells). Our method is based on genetic algorithms and support vector machines: the effectiveness of a subset of expressed genes is evaluated in terms of its discrimination power by SVMs; GA is applied to search in the combinational space of feature subsets in parallel to identify the best subsets. We now detail our GA/SVM algorithms.

### 2.1 Support Vector Machines (SVM)

Typical microarray data consist of tens of data points with thousands or tens of thousands of genes. The limited number of training samples together with high dimensionality of feature space may cause poor generalization problems (e.g., the selected subset of genes may perform well on training samples, but poorly on new samples). We thus consider the use of SVMs to evaluate the effectiveness and discrimination ability of subsets of genes, since SVMs promise good generalization performance for the small sample set limit (Vapnik, 1998).

The foundations of SVMs have been developed by Vapnik (1998). Instead of minimizing the traditional empirical risk (the error on the training data), SVMs minimize an upper bound on the expected risk derived from the capacity of hypotheses. This is made possible by constructing a classifier that separates training samples and maximizes the margin (the minimum distance between the decision surface and training samples).

Consider a two-class classification problem, where the training set is described as

$$(y_1, \mathbf{x}_1), \dots, (y_m, \mathbf{x}_m), \quad \mathbf{x}_i \in R^n, y_i \in \{-1, +1\} \quad (1)$$

where  $y_i$  are class labels. SVMs find an optimal hyperplane that produces large margins. The hyperplane is defined by the equation

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \quad (2)$$

where  $\mathbf{w}$  is the  $n$ -dimensional vector perpendicular to the hyperplane and  $b$  is the bias. Maximizing the margin is equivalent to minimizing the weight norm  $\|\mathbf{w}\|^2$  subject to

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \geq 0. \quad (3)$$

By introducing positive Lagrange multipliers  $\alpha_i, i = 1, \dots, m$  for the above constraints, we can form a primal Lagrange function as follows

$$L_p = \|\mathbf{w}\|^2 / 2 - \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1], \quad (4)$$

This is a convex quadratic programming problem. It is equivalent to solve the “dual” problem (Vapnik, 1998): maximize  $L_p$  subject to the constraints that the gradient of  $L_p$  with respect to  $w$  and  $b$  vanishes. The solution to the dual problem is given by solving the following equation

$$\alpha = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^m \alpha_i, \quad (5)$$

with constraints

$$\alpha_i \geq 0, \quad i = 1, \dots, m, \quad (6)$$

$$\sum_{i=1}^m \alpha_i y_i = 0, \quad (7)$$

In the solution, these samples with nonzero  $\alpha_i$  are called support vectors, which are critical samples for classification.

To build a nonlinear decision surface, we can map these training data into a higher dimensional space where a separating hyperplane is found which maximizes the margin. Hyperplanes can be represented in a Hilbert space by means of kernel functions  $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$  (Osuna et al., 1996). The decision function is then

formulated in terms of these kernels as  $f(x) = \text{sign} \left( \sum_{i=1}^p \lambda_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \right)$ , where the coefficients

$\lambda_i$  are the Lagrange multipliers.

In our application to microarray data, we use linear support vector machines to evaluate the fitness of subsets of genes. This is because the number of features (genes) is much larger than the number of training samples and thus, we expect all training samples are linearly separable.

## 2.2 Feature Selection Using Genetic Algorithms/Support Vector Machines

We now detail the use of GAs to identify the best subsets of genes with discrimination power evaluated by linear SVMs.

The genetic algorithm, first introduced by John Holland in 1975, was the method for the search of an optimal set of feature vectors (Holland, 1995). It has the advantage of applying a random search that is also being directed toward the goal of a best subset. The GA begins a search by generating a large array of randomly generated vectors. It evaluates a set of trial solutions in terms of a fitness function and usually selects those with above average performance with high probability. From the chosen trial solutions, some trial solution pairs are selected for the element swapping by an operation called crossover, while other trial solutions have some of their elements perturbed by mutation (Goldberg, 1989). A new generation consisting of potential solutions is created by the above average fitness solutions and some of the less than optimal valued solutions. This creates regions in solution space that are beginning to converge to a solution, while maintaining a search in the less than optimal areas to prevent the GA from selecting a local extreme. This process of evaluating performance, mating, and mutation causes the GA to accumulate those elements of the trial population that converge on a solution, while maintaining some diversity.

We apply GA to identify subsets of genes that can correctly classify all training samples while at the same time are expected to generalize well. To achieve this, SVMs are used to evaluate the fitness of each subset of genes. The schematic diagram of the proposed GA/SVM is shown in Figure 1.

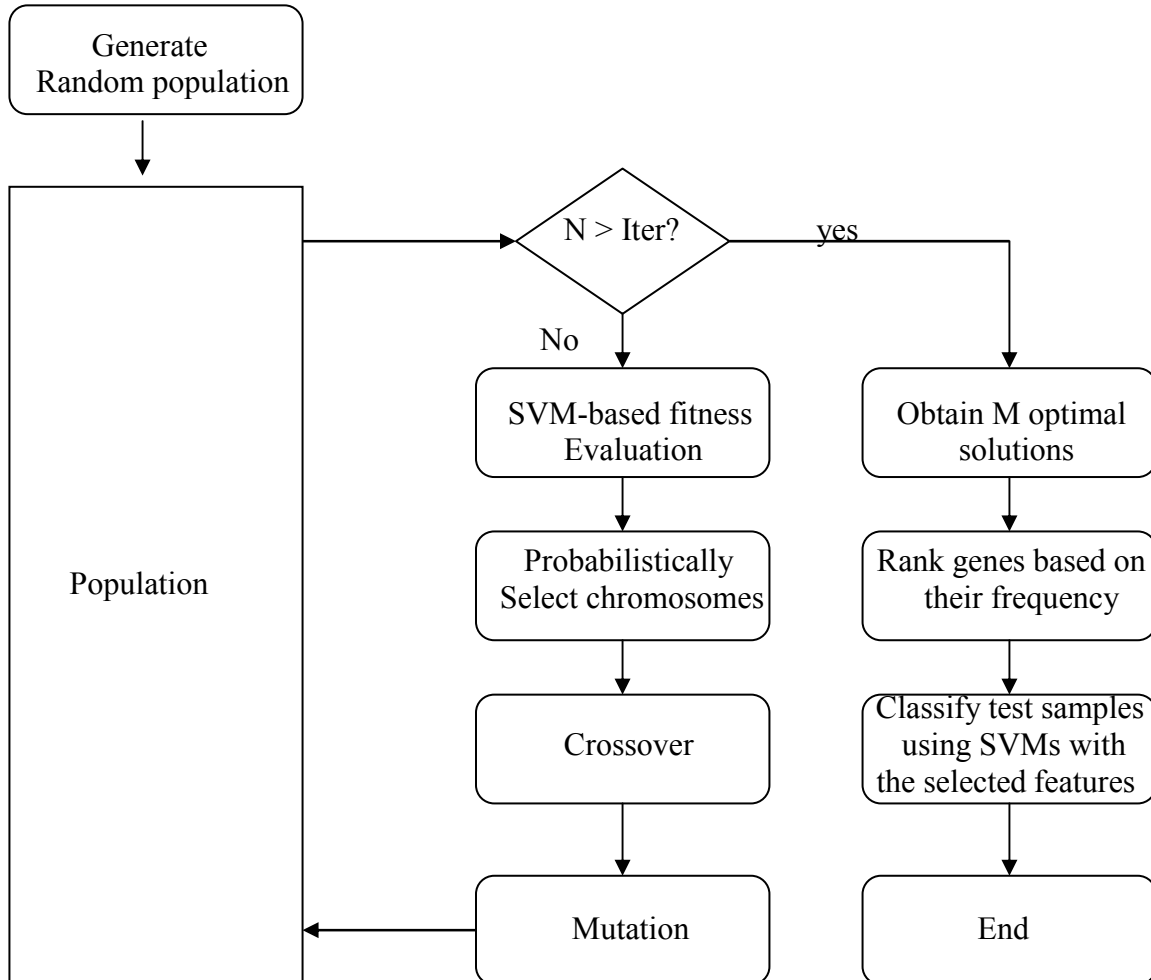


Figure 1: A schematic diagram of the GA/SVM algorithm

*Representation:* To determine the features to use, we will create ‘feature mask’ vectors. Each ‘feature mask’ vector will have a set of indices that address elements in the feature vectors given in the data set.

*Fitness function and evaluation:* An initial population of feature selection masks (chromosomes) is randomly generated with each chromosome consisting of  $d$  distinct genes ( $d$  is the number of features to be selected). The population is evaluated by computing a fitness function: in this study, training data with the subset of features specified by chromosomes are classified by SVMs; the fitness function is then the number of correctly classified training samples using SVMs.

*Selection:* To select chromosomes for next generation, the algorithm calculates the cumulative sum for the fitness of each feature selection mask over the sum of the fitness functions and then applies Roulette wheel selection (Goldberg, 1989). The population for the next iteration is replaced from the current population by random selection. The number of occurrences of a feature selection mask is proportional to its fitness.

*Crossover:* Optimization of the population continues by applying the single point crossover on a subset of the population. The probability of being selected for the subset is dependent upon the ‘Probability of Cross’,  $P_c$  value that is set at the initialization of the algorithm. The subset contains an even number of vectors that are randomly combined by swapping elements from each pair. This promotes the growth of favorable solutions and discourages the growth of the less than favorable solutions. Now, in order to prevent premature convergence, the next operator applied to the population is mutation.

*Mutation:* The mutation operator used in this experiment maintains the number of bits of the feature selection masks at a fixed number  $d$ . We probabilistically select  $l$  chromosomes to mutate: for each of the  $l$  chromosome, randomly select one element and replace this element with an index value (randomly picked) that is not shown in this chromosome.

At this point, the population is replaced and the iteration count has been incremented. When the genetic algorithm has completed all of the specified iterations, the population is again evaluated for fitness and the best feature mask selected and returned.

For microarray data sets, the number of training samples is typically small. Thus, more than one subset of genes that correctly classify training samples may exist. With the GA/SVM methods, we can obtain many such subsets. Similarly to the strategy used in Li et al. (2001), the frequency of each gene selected in these near-optimal solutions is assessed and important genes are expected to have a high frequency to be selected. Because of the SVMs’ generalization ability, we expect the subset of genes selected generalizes well for unseen data.

### **3. EXPERIMENTAL RESULTS**

To evaluate the performance of the proposed GA/SVM algorithm, we used two public microarray datasets for cancer classification: Princeton colon (tumor versus normal) tissue database (Alon et. al., 1999) and MIT leukemia (acute lymphoblastic leukemia (ALL) versus acute myeloid leukemia (AML)) database (Golub et. al., 1999).

#### **Identify genes: Colon cancer dataset**

The colon cancer dataset (Alon et. al., 1999) contains gene expression information extracted from DNA microarrays. This microarray dataset is used to distinguish tumor and normal colon tissues. There are 62 tissue samples, of which 22 are normal and 40 are cancer tissues, each having 2000 genes with highest minimal intensity across the 62 issues. Gene expression levels were measured as ratios of the expression level under a given condition to the expression level under a reference condition. A logarithmic transformation is performed on the expression-ratio, i.e., the features used are the log-ratio of the intensities measured from microarray images. The data set was divided into a training set with 32 samples and a test set with 30 samples.

A total number of 5157 subsets of genes that correctly classify all training samples are obtained using our GA/SVM algorithms. Each subset consists of five genes (i.e.,  $d = 5$ ). Genes are then ordered based on the number of occurrences with which genes are selected. Figure 2

shows the number of occurrences for each gene. As can be seen, some genes are selected significantly more often than other genes. For example, human monocyte-derived neutrophil-activating protein (MONAP) mRNA is selected more than 200 times, while sparc precursor is never selected. To see whether the selection was random, we run the GA/SVM algorithms to identify 4000 subsets and 3000 subsets of five genes and show the corresponding numbers of occurrences in Figures 3 and 4, respectively. It is clear that the top best genes are identical.

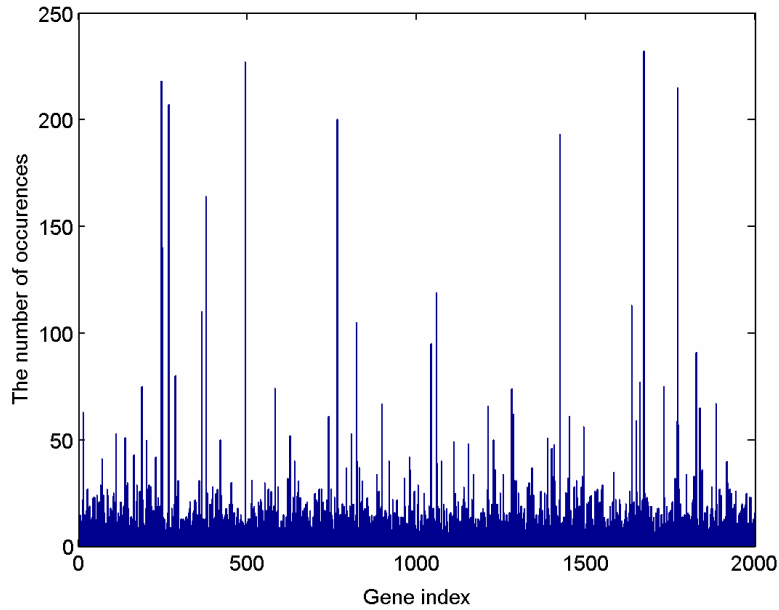


Figure 2: Genes and the number of their occurrences. 5157 subsets of five genes are identified using GA/SVM algorithms. Each subset can correctly classify all training samples.

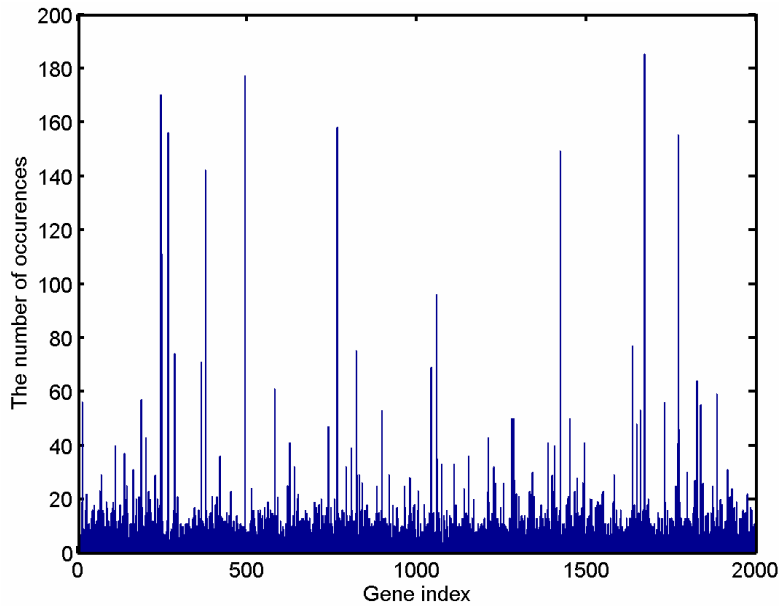


Figure 3: Genes and the number of their occurrences. 4000 subsets of five genes are identified using GA/SVM algorithms. Each subset can correctly classify all training samples.

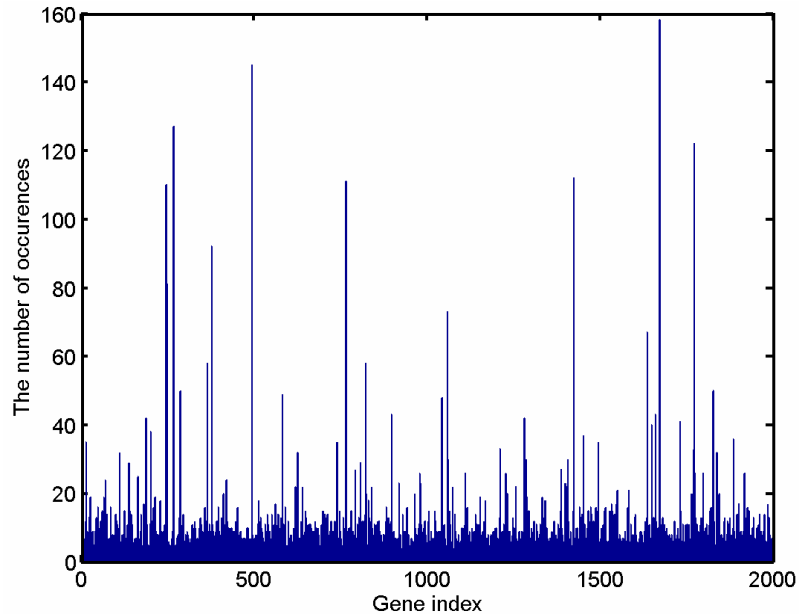


Figure 4: Genes and the number of their occurrences. 3000 subsets of five genes are identified using GA/SVM algorithms. Each subset can correctly classify all training samples.

The top 25 genes ranked by the number of occurrences using the GA/SVMs are listed in Table 1. The most frequently selected gene is MONAP mRNA, whose expression level is directly related to tumor angiogenesis and aggression (Xu et al., 1999). Three muscle genes (Myosin regulatory light chain 2 (smooth muscle isoform), Tropomyosin, fibroblast and epithelial muscle-type, and Myosin light chain alkali (smooth-muscle isoform)) are also selected. It is believed that normal tissues and tumors have different muscle indexes (Alon et al., 1999). Notterman et al. (2001) have demonstrated that in early colon tumors, uroguanylin is significantly reduced. Thus, it is not surprising that H.sapiens mRNA for GCAP-II/uroguanylin precursor is selected. Another frequently selected gene is the vasoactive intestinal peptide (VIP) gene, which plays an important role in the proliferation of various cancer cells (Hilairret et al., 1998, Maruno et al., 1998). In addition, other interesting genes (antigen genes, nonmuscle genes etc.) have also been frequently identified.

To test the discrimination ability of the genes selected by GA/SVM algorithms, we classify the test samples using the top 25 genes. The number of errors for both training and test data set is listed in Table 2. For comparison, Table 2 also includes results obtained with the combined forward selection and k nearest neighbor (FS/kNN) methods and the combined individual ranking and k nearest neighbor (IR/kNN) methods. All methods used are called wrapper methods (the final classifier is used to measure the performance of the selected subset of features). We use the same parameter  $k = 3$  in kNN as Li et al. (2001). With the IR/kNN-selected genes, six training samples are misclassified and 20 out of 30 test samples are misclassified. The FS/kNN algorithms produce better results than IR/kNN for training data set: all training samples are correctly classified. However, 20 out of 30 test samples are misclassified. Among the three algorithms, the proposed GA/SVM algorithm yields the best results. With the 25 GA/SVM-selected genes, all training samples are correctly classified and only six out of 30 test samples are misclassified. This indicates that our GA/SVM algorithm is able to identify regulated genes that, when combined together, can discriminate cancer cells from normal cells. Only few genes are

commonly found by all three algorithms. Most genes identified by the GA/SVM algorithms may not be considered most discriminative features by individual ranking methods when they are measured individually. Together, however, these features may be most discriminative for cancer classification.

Table 1: The 25 best genes most frequently selected by GA/SVMs using training samples

Gene Number	Gene Name
M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA
R87126	Myosin heavy chain, nonmuscle (Gallus gallus)
J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete
M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6
J02854	Myosin regulatory light chain 2, smooth muscle isoform (human)
Z50753	H.sapiens mRNA for GCAP-II/uroguanylin precursor
M63391	Human desmin gene
M80815	H.sapiens a-L-fucosidase gene, exon 7 and 8
M36634	Human vasoactive intestinal peptide (VIP) mRNA
X14958	Human hmgI mRNA for high mobility group protein Y
T92451	Tropomyosin, fibroblast and epithelial muscle-type (human)
R36977	Transcription factor IIIA
M55265	Human casein kinase II alpha subunit mRNA
H64489	Leukocyte antigen CD37 (Homo sapiens)
H87135	Immediate-early protein IE180 (Pseudorabies virus)
X54942	H.sapiens cks2 mRNA for Cks1 protein homologue
T51023	Heat shock protein HSP 90-beta (human)
X16354	Human mRNA for transmembrane carcinoembryonic antigen BGP $\alpha$
T51571	P24480 calgizzarin
R44301	Mineralocorticoid receptor (Homo sapiens)
H43887	Complement factor D precursor (Homo sapiens)
R55310	Mitochondrial processing peptidase
U14631	Human 11 beta-hydroxysteroid dehydrogenase type II mRNA
H20709	Myosin light chain alkali, smooth-muscle isoform (human)
D16294	Human mRNA for mitochondrial 3-oxoacyl-CoA thiolase

Table 2: Performance of classification for GA/SVM, FS/kNN, and IR/kNN methods

Methods	GA/SVM	FS/kNN	IR/kNN
Training errors	0	0	6
Test errors	6	20	20

## Identify genes: Leukemia dataset

The Leukemia dataset (Golub et. al., 1999) consists of 72 samples, of which 47 are ALL and 25 are AML. The gene expression levels of all samples were extracted from microarray images. Each sample has 7129 features (i.e., 7129 genes). We use similar procedures as described in Dudoit et. al. (2002) for data preprocessing: genes with smallest expression level less than 1 and genes with largest expression level greater than 13550 were excluded. A dataset of 1800 genes is available after the preprocessing. A logarithmic transformation is then performed on the 1800 expression levels. The data set was divided into a training set with 38 samples and a test set with 34 samples.

A total number of 2368 subsets of five genes that correctly classify all training samples were obtained using our GA/SVM algorithms. Genes are then ranked by the number of occurrences with which genes were selected. Figure 5 shows the number of occurrences for each gene. Similarly with colon cancer database, some genes were selected significantly more often than other genes.

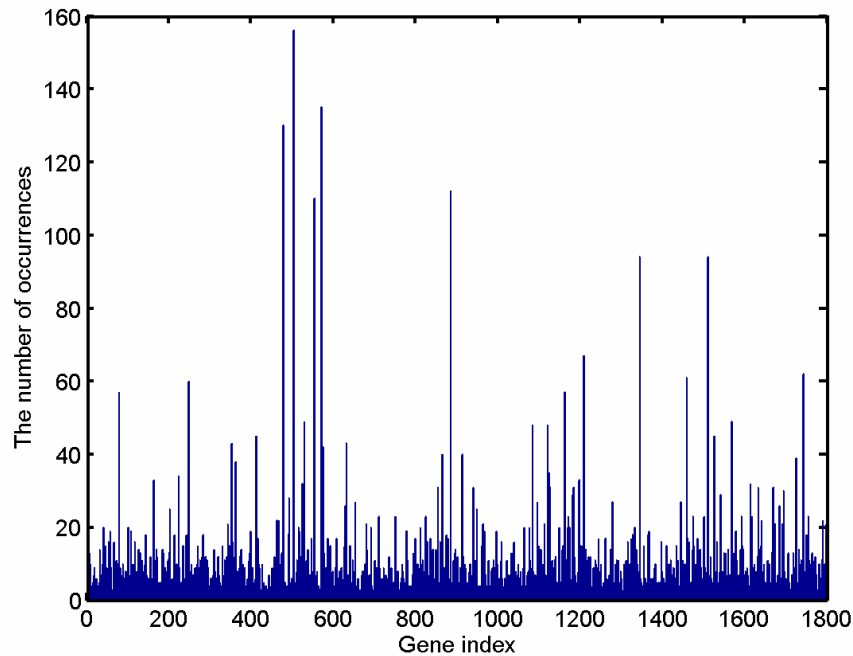


Figure 5: Genes and the number of their occurrences. 2368 subsets of five genes are selected using GA/SVM algorithms. Each subset can correctly classify all training samples.

To evaluate the performance of the selected genes in cancer classification, we classify the test samples (ALL versus AML) using the top 25 genes. The number of errors for both training and test data set is listed in Table 3. Again, we can see that features selected by individual ranking methods perform poorly: six out of 38 training samples and 14 out of 34 test samples are misclassified. Forward selection performs better than individual ranking. Features selected from our GA/SVMs correctly classify all training samples and only misclassify one test sample. Thus, the 25 GA/SVM-selected features are informative for distinguishing ALL from AML.

Table 3: Performance of classification for GA/SVM, FS/kNN, and IR/kNN methods

Methods	GA/SVM	FS/kNN	IR/kNN
Training errors	0	0	6
Test errors	1	7	14

#### 4. CONCLUSIONS

Recently developed microarray technologies allow for monitoring the expression level of thousands of genes in parallel. The processing and exploitation of useful information from gene data sets pose a challenging problem. In this paper, we propose a practical and efficient feature selection algorithm to select informative genes from very high-dimensional spaces. We perform the GA based gene selection in the context of SVMs. Thus, the selected genes are expected to generalize well. This is particularly important in microarray data classifications, since only very limited training samples are available. Secondly, we present a different strategy using GAs. Both crossover and mutation will be operated with modifications in GAs. Two databases for cancer classification are considered: the colon cancer database and the leukemia database. For both datasets, the proposed GA/SVM methods successfully identify genes informative for cancer classification and yield much better results than both forward selection and individual ranking methods. The proposal GA/SVM algorithms are well suited for selecting features from very large feature spaces due to the parallel searching strategies and the ability to generalize well.

#### REFERENCES

- Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999) Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, **96**, 6745-6750.
- Bo, T. and Jonassan, I. (2002) New feature subset selection procedures for classification of expression profiles, *Genome Biology*, **3**, 0017.1-0017.11.
- Cho, R., Campbell, J., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., and Lockart, D. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell*, **2**, 65-73.
- Chu, S., Derisi, J., Eisen, M., Mullholland, J., Botstein, D., Brown, P., and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast, *Science*, **282**, 699-705.
- Cover, T. and Campenhout, J. (1977) On the possible orderings in the measurement selection problem, *IEEE Trans. Systems, Man, and Cybernetics*, **SMC-7(9)**, 657-661.
- De Risi, J., Iyer, V., and Brown, 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. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Amer. Stat. Assoc.*, in press.
- Eisen, M. and Brown, P. (1999) DNA arrays for analysis of gene expression, *Methods in Enzymology*, **303**, 179-205.

- Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998) Clustering analysis and display of genome wide expression patterns, *Proc. Natl. Acad. Sci., USA*, **95**, 14863-14868.
- Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., and Haussler, D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**, 906-914.
- Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization & Machine Learning*, Addison Wesley.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression, *Science*, **286**, 531-537.
- Hilairret, S., Janet, T., Pineau, N., Caigneaux, E., Chadeneau, C., Muller, J., and Philippe, M. (1998) The small G-proteins Rap1 as potential targets of vasoactive intestinal peptide effects in the human colonic cancer cells HT29, *Neuropeptides*, vol. 32 (6), pp. 587-595.
- Holland, J. (1995). *Hidden Order, How Adaptation Builds Complexity*, Perseus Books.
- Inza, I., Sierra, B., Blanco, R., and Lerranaga, P. (2002) Gene selection by sequential search wrapper approaches in microarray cancer class prediction, *Journal of Intelligent and Fuzzy Systems*, in press.
- Kerr, M., Martin, M., and Churchill, G. (2000) Analysis of variance for gene expression microarray data, *J. Computat. Biology*, **7**, 819-825.
- Li, L., Darden, T., Weinberg, C., Levine, A., and Pederson, L. (2001) Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. *Combinational Chemistry and High Throughput Screening*, vol. 4(8), pp. 727-739.
- Long, A., Mangalam, H., Chan, B., Toller, L., Hatfield, G., and Baldi, P. (2001) Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework, *J. Biol. Chem.*, **276**, 19937-19944.
- Maruno, K., Absood, A., and Said, S. (1998) Vasoactive intestinal peptide inhibits human small-cell lung cancer proliferation *in vitro* and *in vivo*. *Proc. Natl. Acad. Sci.* vol. 95, pp.14373-14378
- Newton, M., Kendzioriski, C., Richmond, C., Blattner, F., and Tsui, K. (2001) On differential variability of expression ratios: improved statistical inference about gene expression changes from microarray data, *J. Comput. Biol.*, **8**, 37-52.
- Notterman, D., Alon, U., Sierk, A., and Levine, A. (2001) Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.*, vol. 61, pp. 3124-3130.
- Osuna, E., Freund, R., and Girosi, F. (1996) *Support vector machines: Training and applications*. Technical Report AIM-1602, MIT A.I. Lab.
- Raymer, M., Punch, W., Goodman, E., Kuhn, L., and Jain, A. (2000) Dimensionality reduction using genetic algorithms, *IEEE Trans. on Evolutionary Computation*, vol. 4(2), pp. 164-171.
- Vapnik, V. (1998) *Statistical Learning Theory*. Wiley, New York.
- Xing, E., Jordan, M., and Karp, R. (2001) Feature selection for high-dimensional genomic microarray data, In *Proceedings of Eighteenth International Conference on Machine Learning*, San Francisco.
- Xu, L., Xie, K., Mukaida, N., Matsushima, K., and Fidler, I. (1999) Hypoxia-induced elevation in interleukin-8 expression by human ovarian carcinoma cells. *Cancer Res.*, vol. 59, pp.5822-5829.