# GENETIC ALGORITHMS AND SILHOUETTE MEASURES APPLIED TO MICROARRAY DATA CLASSIFICATION

TSUN-CHEN LIN, RU-SHENG LIU, SHU-YUAN CHEN

*Dept. of Computer Science and Engineering, Yuan Ze University*
*135 Yuan-Tung Rd, Nei-Li, Chung-Li, Taoyuan, 32026, Taiwan*

CHEN-CHUNG LIU

*Graduate Institute of Network Learning Technology, National Central University*
*300 Jhongda Rd, Jhongli City, Taoyuan, 320, Taiwan*

CHIEH-YU CHEN

*Graduate School of Biotechnology and Bioinformatics, Yuan Ze University*
*135 Yuan-Tung Rd, Nei-Li, Chung-Li, Taoyuan, 32026, Taiwan*

Microarray technology allows large-scale parallel measurements of the expression of many thousands genes and thus aiding in the development of efficient cancer diagnosis and classification platforms. In this paper, we apply the genetic algorithm and the silhouette statistic in conjunction with several distance functions to the problem of multi-class prediction. We examine two widely used sets of gene expression data, measured across sets of tumors, and present the results of classification accuracy on these two datasets by our methods. Our best success rate of tumor classification has better accuracy than many previously reported methods and it provides a useful method towards a complete tool in this domain.

## 1. Introduction

Microarray technology allows large-scale parallel measurements for the expression of many thousands of genes and produces a very large amount of gene data. One of the most promising applications of this technology is as a useful tool for tumor classification and cancer diagnosis, and several analytical approaches have been applied for this task such as Golub *et al.*,[3] Ben-Dor *et al.*,[4] Alizadeh *et al.*[5] Currently, these reported techniques have focused on the problem where the expression profiles which contain only two or three classes and gave the results, returning test success rates close to 90-100% for most of the binary class data. However, when this problem of tumor classification is expanded to multiple tumor classes, the performance of these methods decreases significantly because gene expression hallmarks of classification for different cancer types are still not clearly defined.[9] This makes it for methods like Golub *et al.*[3] or Slonim *et al.,*[6] based on gene expression, starting with a feature selection to take possible correlation with an ideal gene marker particularly difficult. Furthermore, due to the complex relationships among the genes that may affect the discriminate analysis in classification, there is still less attention paid to the discriminate approaches, which consider take the co-action among the genes.

Two more recent approaches have addressed this problem and applied linear discriminate analysis with Mahalanobis metric to compute the distance to centroids and Quadratics discriminate analysis with genetic algorithms (GAs) used for feature selection.[11, 12] This leads to our proposal, which is the use of GAs to identify a set of key features and combine the silhouette statistic with a form of linear discriminate analysis. The key issue of this paper is in the context of using GA/silhouette methods not only to identify a set of key genes but also to take the co-action factors among these genes into consideration. This is in contrast to other methods, which only make comparisons between pair of genes (gene vs. ideal gene), and thus why they may not produce the comparable accuracies for more complex datasets. To achieve the maximum discriminate ability for genes to classify tumor samples, we have also explored several distance metrics to evaluate their sensitivities for the discriminate analysis. Where GAs were first used to summarize the input spaces into an selected subspace and evolved the selections to the optimal space by measuring the silhouette statistic with the one-minus-Pearson distance metric, our methodologies finally exhibited a 4% improvement in classification accuracies over several recently reported techniques, based on the same experimental dataset.

## 2. Methods

### 2.1. *The Silhouette Statistic*

The silhouette statistic of Kaufman and Rousseeuw has been used to study the quality of clustering by the method that measures how well an item is assigned to its corresponding cluster.[2] In this section, we extend this concept to describe the main discriminate method that we will use in this paper. Our algorithm starts by assuming that we are given training set *D*. Let $D = \{\langle \vec{e}_i , l_i \rangle$, for $i$=1…m} be a set of *m* training samples, where $\vec{e}_i = (e_{i1}, e_{i2}, ..., e_{iG})^T$ is the vector of $i^{\text{th}}$ sample in $R^G$ that describes expression levels of *G* predictive genes, and $l_i \in L = \{1,2\ldots q\}$ is the class label associated with $\vec{e}_i$. Our discriminate function based on the silhouette statistic is then defined as

$$s(\vec{e}_i) = \frac{b(\vec{e}_i) - a(\vec{e}_i)}{\max\{a(\vec{e}_i), b(\vec{e}_i)\}}$$

In our definition, $d(\vec{e}_{i,} c_s)$ denotes the average distance of $i^{\text{th}}$ sample to other samples in the class of $c_s$, $b(\vec{e}_i)$ denotes $\min\{d(\vec{e}_{i,} c_s)\}$, $\vec{e}_i \in c_r$, $r \neq s$, $r \in \{1,2\ldots q\}$, q is the number of classes, and $a(\vec{e}_i)$ denotes $d(\vec{e}_{i,} c_s)$, $\vec{e}_i \in c_r$, $r = s$. The $s(\vec{e})$, ranging from -1 to +1, is the discriminate function, returning the score to indicate how well an input sample can be assigned to its own class under the vector of $\vec{e}$ . For example, in a domain of q classes, a predictive gene set chosen from a selection method will constitute $\vec{e}_i$ for the $i^{\text{th}}$ sample and used for calculating the silhouette value (discriminate score) to decide how well this set of genes represents the sample associated with its class. Essentially, this function uses the ratio of between-groups variance to within-groups variance in order to measure the

$s(\vec{e}_i)$ value and determines whether the associated class label $l_i$ is the predicted label of the query sample $\vec{e}_i$. In our algorithm throughout the experiments, we set the threshold to $s(\vec{e}) \geq 0$ for all samples in the dataset. In words, once the returning value is less than zero, we say the corresponding sample is misclassified under the discriminate variable of $\vec{e}$. Therefore, the classification rule can be written as

$$\mathrm{C}(\vec{e}_i) = l_i, \text{ where } s(\vec{e}_i) \geq 0.$$

### 2.2. *Genetic Algorithms*

To classify samples using microarrays, it is necessary to decide which genes should be included to form the sample vector (predictor set). Gene selections here for the classifications of multiple cancer types were based on a group of genes chosen by GAs and used by the discriminate analysis. The genetic algorithms were provided originally from the report of Ooi and Tan,[12] with toolboxes of two selection methods: (1) stochastic universal sampling (SUS) and (2) roulette wheel selection (RWS). In addition two tuning parameters, *Pc*: crossover rate and *Pm*: mutation rate, were used to tune one-point and uniform two crossover operations in order to evolve the population of individuals for the mating pool.

To determine the fitness and find increasing fit combinations of predictor genes in a chromosome represented by strings $Si$, $Si = [R\ g_1\ g_2\ \ldots\ g_{i=Rmax}]$, the GA method defined the fitness function of $f(Si) = 200 - (E_C + E_I)$, where $R$ denotes the size of the predict set of genes, $E_C$ is the error rate of leave-one-out cross-validation (LOOCV) test on the training data, and $E_I$ is the error rate of independent test on the test data. The genes in the string $S_i$, are used to form a sample vector representing sample's expression and are evaluated by our discriminate function to classify tumor samples. Intuitively, the silhouette statistic $\mathrm{s}(\vec{e})$ combining the selection methods of GAs then will find the best quality of $\vec{e}$ to discriminate samples between two or more existing groups.

### 2.3. *Running the GA/silhouette Algorithm and Estimating Prediction Errors*

The running of our approach begins with setting 100 individual runs to the GA/silhouett*e* algorithm, with each run beginning with a different initial gene pool in order to have an unbiased estimation of classifier performance. The maximum generations for GAs are set to 100, for which each generation produces 100 and 30 chromosomes, containing the size of genes ranging from 11 to 15 and 5 to 50, in a chromosome corresponding to the NCI60 and the GCM dataset respectively. In addition, the following procedures are used to examine the assessment of accuracy for each chromosome in predicting the class of an unknown sample $\vec{e}_i$.

```
1.  FOR each chromosome $C_i$ to $\mathrm{C}_{max}$
2.  FOR each leave-one-out training sample $\vec{e}_i \in l_i$
```

```
IF ($s(\vec{e}_i)$<0)
$Xc$Error = $Xc$Error + 1    // Sample misclassified
END FOR
FOR each independently unknown samples $\vec{e}_i \in l_i$
IF ($s(\vec{e}_i)$<0)
$Xi$Error = $Xi$Error + 1    // Sample misclassified
END FOR
END FOR
$Ec$ErrorRate = $Xc$Error / Total training samples   //  LOOCV error rate
$Ei$ErrorRate = $Xi$Error / Total test samples        //  Independent test error rate
Fitness  = 200 – $Ec$ErrorRate + $Ei$ErrorRate        //  Fitness of  a chromosome
END FOR
Findmax (Fitness )
Go next generation
```

### 2.4. *Distance Matrices*

Our discriminate method is a function depending on two arguments, the distance function, and a query $\vec{e}_i$ . As generally known, the distance metrics that are used generally have a large effect on the performance of the discriminant analysis. Therefore, we applied several distance functions as described in Table 1 to deal with the silhouette statistic $s(\vec{e}\,)$ and explore the best discriminate ability for genes in classification.

**Table 1.** Distance metrics and Formula (Source Dudoit and Fridlyand.[8]).

| Name | Formula |
|---|---|
| Euclidean metric | $d_E(\vec{e}_{i,}\vec{e}_j) = \{\Sigma_G(\vec{e}_{Gi} - \vec{e}_{Gj})^2\}^{1/2}$ |
| Mahalanobis metric | $d_{Ml}(\vec{e}_{i,}\vec{\mu}_j) = (\vec{e}_i - \vec{\mu}_j)S^{-1}(\vec{e}_i - \vec{\mu}_j)'$<br>$S$ is the between classes common covariance matrix.[1]<br>$\vec{\mu}_j$ is the class mean vector of class j. |
| Manhattan metric | $d_{Mn}(\vec{e}_{i,}\vec{e}_j) = \Sigma_G\left|\vec{e}_{Gi} - \vec{e}_{Gj}\right|$ |
| Canberra metric | $d_c(\vec{e}_{i,}\vec{e}_j) = \Sigma_G\left|(\vec{e}_{Gi} - \vec{e}_{Gj})/(\vec{e}_{Gi} + \vec{e}_{Gj})\right|$ |
| One minus Pearson metric | $d_P(\vec{e}_i,\vec{e}_j) = 1 - \dfrac{\Sigma_G(\vec{e}_{Gi} - \vec{e}_{\bullet i})(\vec{e}_{Gj} - \vec{e}_{\bullet j})}{\{\Sigma_G(\vec{e}_{Gi} - \bar{\vec{e}}_{\bullet i})^2\}^{1/2}\{\Sigma_G(\vec{e}_{Gj} - \bar{\vec{e}}_{\bullet j})^2\}^{1/2}}$ |

## 3. Datasets

There are two published microarray datasets from human cancer cell lines used in this paper. Before the datasets were used in our experiments, the data was preprocessed by following steps.

1. The spots with missing data, control, and empty spots were excluded.

2. For each sample array in both datasets, the Cy5/Cy3 ratio of every spot is normalized by subtracting the mean of Cy5/Cy3 ratio of control spots and dividing the result by the standard deviation of control spots.
3. A preliminary selection of 1000 genes with the highest ratios of their between-groups to within–groups sum of squares (BSS/WSS) was performed. For gene *i,* $x_{ij}$ denotes the expression level from patient *j*, and the ratio is defined as

$$\frac{BSS(i)}{WSS(i)} = \frac{\sum_{j=1}^{M_t}\sum_{q=1}^{Q} I(c_j = q)(\mu_{qi} - \mu_{\bullet i})^2}{\sum_{j=1}^{M_t}\sum_{q=1}^{Q} I(c_j = q)(x_{ij} - \mu_{qi})^2}$$

where $M_t$ is the training sample size; $Q$ is the number of classes; and $I(\bullet)$ is the indicator function which equals 1 if the argument inside the parentheses is true, and 0 otherwise. Furthermore $\mu_{\bullet i}$ denotes the average expression level of gene *i* across all samples, and $\mu_{qi}$ denotes the average expression level of gene *i* across all samples belonging to class *q*. This is the same calculation used in Dudoit *et al.*[8] In our cases, the BSS/WSS ratios for NCI60 data ranges from 0.4 to 2.613, and from 0.977 to 3.809 for GCM data.

### 3.1. *The NCI60 Dataset, Ross et al.*[7]

The NCI60 gene expression levels were measured with 9,703 spotted cDNA sequences among the 64 cell lines from tumors with 9 different sites of origin from the National Cancer Institute's anti-cancer drug screen and can be downloaded from http://genome-www.stanford.edu/sutech/download/nci60/dross_array_nci60.tgz. During the data preprocessing, the single unknown cell line and two prostate cell lines were excluded due to their small sample size, leaving a matrix of 1000 genes × 61 samples. These genes are henceafter referred to by their index numbers (1 to 1000) in our experiments. To build the classifier, this dataset was divided into the ratio as 2:1 (41 samples for training and 20 for testing). The 41 patient samples are gene expression levels composed of 5 breast, 4 central nervous system (CNS), 5 colon, 4 leukemia, 5 melanoma, 6 non-small-cell-lung-carcinoma (NSCLC), 4 ovarian, 5 renal, and 3 reproductive.

### 3.2. *The GCM Dataset, Ramaswamy et al.*[10]

The gene expression levels were measured by Affymetrix Genechips containing 16063 genes among 198 tumors with 14 different classes of tumor, and can be downloaded from http://www-genome.wi.mit.edu/mpr/publications/projects/Global_Cancer_Map/. During data preprocessing, the dataset left a matrix of 1000 genes × 198 samples. These genes are referred to by their index numbers (1 to 1000) in our experiments. This dataset originally contained 144 samples for training, and 54 for test. The 144 patient samples are gene expression levels composed of 8 breast, 8 prostate, 8 lung, 8 colorectal, 16 lymphoma, 8 bladder, 8 melanoma, 8 uterine, 24 leukemia, 8 renal, 8 pancreatic, 8 ovarian, 8 mesothelioma, and 8 brain.

## 4. Results and Discussions

The Tables 2 and 3 summarize the parameters, gene selection and crossover operation, and distance metrics used in the GA/silhouette algorithm and the results corresponding to the LOOCV test on the training data and independent blind tests on the test data. After finishing the 100 runs, the best chromosome of populations with the best fitness, chosen from the simulations to arrive at the optimal operation is based on the idea that a classifier need not only to work well on the training samples, but also must work equally well on previously unseen samples. Therefore, the optimal individuals of each generation were sorted in ascending order by the sum of the error number on both tests. The smallest number then decides the chromosome that contains the optimal predictor set of genes and gives the number of test errors obtained from our methods.

### 4.1. *GA Parameters and Classification Accuracies of the NCI60 Dataset*

Following the parameters that can produce the best predictor gene sets, as described in Ooi and Tan,[12] we set *Pc* and *Pm* equally to make the performances easy to compare among different distance matrices and described the results in Table 2.

**Table 2.** *Xc*: Cross-validation errors (41 samples); *Xi*: independent test errors (20 samples); *Xa*: total errors (*Xc* + *Xi*); and *R*: the number of predictive genes.

| | | | | 1-Pearson | | | | Mahalanobis | | | | Euclidean | | | | Manhattan | | | | Canberra | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pc | Pm | Crossover | Selection | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* |
| 1 | 0.002 | Uniform | SUS | 4 | 2 | 6 | 14 | 7 | 1 | 8 | 14 | 10 | 4 | 14 | 15 | 11 | 2 | 13 | 15 | 9 | 6 | 15 | 15 |
| 0.7 | 0.005 | One-point | SUS | 5 | 2 | 7 | 15 | 7 | 2 | 9 | 13 | 11 | 4 | 15 | 14 | 14 | 4 | 18 | 13 | 12 | 6 | 18 | 12 |
| 0.7 | 0.001 | Uniform | RWS | 6 | 3 | 9 | 15 | 11 | 2 | 13 | 14 | 13 | 4 | 17 | 13 | 12 | 5 | 17 | 14 | 11 | 8 | 19 | 14 |
| 0.8 | 0.02 | One-point | RWS | 6 | 1 | 7 | 15 | 9 | 3 | 12 | 13 | 15 | 1 | 16 | 15 | 12 | 4 | 16 | 14 | 11 | 8 | 19 | 15 |

The results listed in the Table 2 show the GA/silhouette method with the one-minus-Pearson metric, the SUS and Uniform strategy, indicating that the 14 predictor genes can achieve the best cross-validation training error rate equal to 9.75% (4 errors out of 41samples) and the best test error rate equal 10% (2 errors out of 20 samples). The average performances of the second best are the use with the Mahalanobis metric with 13 to 14 features, while the worst case is the Canberra metric. We also found that when one method outperformed the other in one set of parameters it would also trend to perform better than other sets of parameters.

### 4.2. *GA Parameters and Classification Accuracies of the GCM Dataset*

Having obtained good performance on the NCI60 dataset with 9 classes, we next tested the proposed method on a more complicated dataset consisting of 14 classes, with each class containing more samples to examine the generality of this method. By using our experience with the NCI60 dataset as a guide to select appropriate parameters for the GA/silhouette in the classification to the GCM dataset, we utilized the uniform with SUS strategy and tried two cases by setting the *Pc*=0.98, 0.8 and *Pm*=0.002, 0.001 respectively. As shown in Table 3, the outcomes of the one-minus-Pearson metric were still better than the other metrics. Therefore, we took our concerns about the gene selection/sample classification with the choices of distance metrics to compare the sensitivities among different distance metrics by taking GCM data, for example in Figure 1.

**Table 3.** *Xc*: Cross-validation errors (144 samples); *Xi*: independent test errors (54 samples); *Xa*: total errors (*Xc* + *Xi*), and *R*: the number of predictive genes.

| | | | | 1-Pearson | | | | Mahalanobis | | | | Euclidean | | | | Manhattan | | | | Canberra | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pc | Pm | Crossover | Selection | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* | *Xc* | *Xi* | *Xa* | *R* |
| 0.98 | 0.002 | Uniform | SUS | 22 | 9 | 31 | 50 | 29 | 6 | 35 | 32 | 46 | 21 | 67 | 43 | 51 | 22 | 73 | 36 | 69 | 26 | 95 | 26 |
| 08 | 0.001 | Uniform | SUS | 28 | 8 | 36 | 50 | 28 | 10 | 38 | 35 | 50 | 18 | 68 | 50 | 52 | 23 | 75 | 38 | 64 | 32 | 96 | 32 |

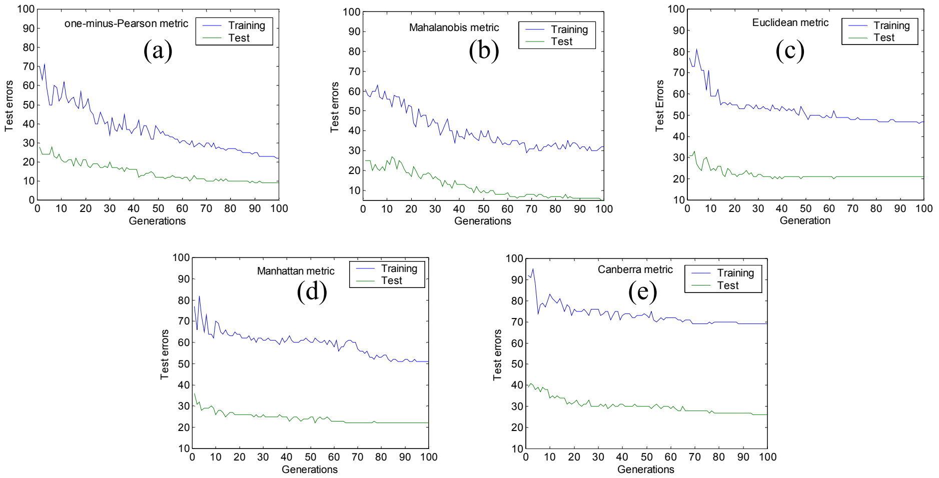

Figure 1. The number of training set (top line) and test set (bottom line) errors that were obtained based on Pc=0.98, Pm=0.002 with the Uniform and SUS strategy from the best run out of 100 individual runs.

For the outcomes of classification accuracies on the NCI60 and GCM datasets, and the learning processes on Figure 1(a) and Figure 1(b), we can see that the GA/silhouette algorithm with one-minus Pearson metric has the comparable performances to Mahalanobis metric whereas the Pearson correlation distance metric seems to have a slightly better discrimination on the LOOCV test for training samples. Furthermore, the learning trends of different metrics shown in the Figure 1 also clearly illustrate the functionalities of these metrics.

### 4.3. *Comparisons of Classification Accuracies with Other Methodologies*

Since the NCI60 and the GCM datasets have become two popular benchmark datasets used by many classification algorithms, we list the classification accuracies of some previous published methods for comparison with ours.

From Table 4, we can see the GA/silhouette method outperforms many previous published methods on the NCI60 data. The best results achieve 90% accuracy (4 errors out of 41samples) in the LOOCV test on the training data and 90% accuracy (2 errors out of 20 samples) in independently blind tests on the test data.

For the GCM dataset, it was reported that this dataset could be classified into their classes by methods returning promising test errors.[10, 12, 13] In our comparisons with these methods, the best model of our methods yielded clear improvement over the others listed in Table 5. Even while the best classification accuracy from our test provided performance equal to the GA/SVM/RFE method, based on the same crossover rate = 0.98 and mutation rate = 0.002, this method took all samples for training (leaving one sample for testing and training with the remaining 197 samples) and made the LOOCV test only, when we only used 144 samples for LOOCV training and tested with 54 samples through training models.

**Table 4.** Recognition success rate comparison with some other algorithms for the NCI60 dataset.

| Classification Method | LOOCV (%) | Independent test (%) | Overall (%) | No. of genes |
|---|---|---|---|---|
| GA/silhouette | 90.3 | 90 | 90.2 | 14 |
| Hierarchical clustering[7] | 81 | - | - | 6831 |
| GA/MLHD[12] | 83 | 95 | 89 | 13 |
| GA/SVM/RFE[13] | 87.93 | - | - | 27 |

**Table 5.** Recognition success rate comparison with some other algorithms for the GCM dataset.

| Classification Method | LOOCV (%) | Independent test (%) | Overall (%) | No. of genes |
|---|---|---|---|---|
| GA/silhouette | 85 | 83 | 84 | 50 |
| GA/SVM/RFE[13] | 85.19 | - | - | 26 |
| GA/MLHD[12] | 79.33 | 86 | 82 | 32 |
| OVA/SVM[10] | 81.25 | 78.26 | 79.75 | 16063 |
| OVA/KNN[10] | 72.92 | 54.34 | 63.63 | 100 |

## 5. Conclusions

Since microarray data analysis has some similarity with information theory, a machine learning approach that discovers subtle pattern in the data is required. Our results indicate that the GA/silhouette algorithm with the one-minus-Pearson distance metric achieved the best performance and outperformed many previous methods. Clear improvements were found with our approach, where the designed classification model gave a simple method and found gene expression fingerprints to allow accurate classification.

In the multi-class classification scenario, the currently available datasets contain relatively few samples but a large number of variables, thus making it difficult to demonstrate one method's superiority. Although no method have yet become a standard to be adopted in this domain, we have demonstrated the use of the GA/silhouette method, which uses the optimal subspace of genes in microarray to take advantage of considering co-relation through the one-minus-Pearson metric among predictor genes in order to find the most important genes to classify samples into a group. From the success of the proposed method, we hope that the use of the GA/silhouette method would be a helpful tool leading to practical uses of microarray data in cancer diagnosis.

**References**

1. M. James. Classification Algorithms. Wiley, New York, 1985.
2. L. Kaufman and P. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.
3. T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286: 531—537, 1999.
4. A. Ben-Dor, N. Friedman and Z. Yakhini. Scoring genes for relevance. Technical Report AGL-2000-13 Agilent Laboratories, 2000.
5. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403: 503—511, 2000.
6. D. Slonim, P. Tamayo, J. Mesirov, T. Golub and E. Lander. Class prediction and discovery using gene expression data. In *Proceeding of the fourth annual international conference on computational molecular biology*, 263—272.
7. D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein and P. O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, 24: 227—235, 2000.
8. S. Dudoit, J. Fridly and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *JASA*. Berkeley Stat. Dept. Technical Report #576, 2000.
9. D. Hanahan and R. Weinberg. The hallmark of cancer. *Cell,* 100: 57—71, 2000.
10. C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. S. Lander, J. P. Mesirov and T. R. Golub. Molecular classification of multiple tumor types. *Bioinformatics,* 17: S316—S322, 2001.
11. R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA.*, 99: 6567—6572, 2002.
12. C. H. Ooi and Patrick. Tan. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics,* 19: 37—44. 2003.
13. S. Peng, Q. Xu, X. B. Ling, X. Peng, W. Du, L. Chen. Molecular classification of

cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters,* 555:358—362, 2003.