Multi-class cancer subtype classification based on gene expression signatures with reliability analysis

Li M. Fu[a,b,*], Casey S. Fu-Liu[b]

[a] *University of Florida, Gainesville, FL, USA*
[b] *Pacific Tuberculosis and Cancer Research Organization, Los Angeles, CA, USA*

Received 9 January 2004; accepted 28 January 2004

First published online 24 February 2004

Edited by Gianni Cesareni

Abstract Differential diagnosis among a group of histologically similar cancers poses a challenging problem in clinical medicine. Constructing a classifier based on gene expression signatures comprising multiple discriminatory molecular markers derived from microarray data analysis is an emerging trend for cancer diagnosis. To identify the best genes for classification using a small number of samples relative to the genome size remains the bottleneck of this approach, despite its promise. We have devised a new method of gene selection with reliability analysis, and demonstrated that this method can identify a more compact set of genes than other methods for constructing a classifier with optimum predictive performance for both small round blue cell tumors and leukemia. High consensus between our result and the results produced by methods based on artificial neural networks and statistical techniques confers additional evidence of the validity of our method. This study suggests a way for implementing a reliable molecular cancer classifier based on gene expression signatures.

Key words: Microarray; Functional genomics; Bioinformatics; Cancer; Classification; Gene expression

1. Introduction

Cancers are conventionally classified by the type of tissue in which the cancer originates. However, the subjective interpretation of the histopathology of a cancer specimen is subject to human error or bias, as illustrated by a study in which an agreement rate of only 41% was observed among independent pathologists regarding the subtypes of lung adenocarcinoma [1]. Moreover, many cancers are atypical or lack distinctive morphological features for correct differential diagnosis. To complicate the matter, cancers with similar histopathological appearances may differ substantially in terms of therapeutic responses and clinical courses. For example, it is difficult to make differential diagnosis among the group of small round blue cell tumors (SRBCTs) including four histologically similar types of tumor, but accurate diagnosis is crucial because treatment and prognosis vary depending on the tumor type [2].

The limitation of the morphology-based approach to cancer classification has led to molecular classification, which promises more objective and accurate cancer classification. Techniques such as immunohistochemistry and reverse transcription polymerase chain reaction are used to detect cancer-specific molecular markers, but pathognomonic molecular markers are unfortunately unavailable for most solid tumors [3]. Furthermore, molecular markers do not guarantee a definitive diagnosis owing to possible failure of detection or presence of marker variants.

Microarray-based gene expression profiling has recently emerged as a promising approach to cancer classification for diagnostic, therapeutic, and prognostic decisions. This approach has gained increasing interest, following the success in demonstrating that microarray data differentiated two types of leukemia [4]. DNA microarrays measure gene expression on a genomic scale to determine which genes are active or silent in cancer or normal cells, permitting simultaneous analysis of multiple known or unknown markers. The microarray-based approach has become a modern trend in cancer research and management [4–11].

It is generally more difficult to differentiate subtypes of cancer with similar histological pictures than those with distinctive appearances. Diagnosis involving multiple cancer categories is also more difficult than the case of two categories. Multi-class cancer subtype classification based on statistical techniques [12] and artificial neural networks [2] has been demonstrated.

Microarray data analysis is characterized by extremely high data dimensionality due to a large number of gene expression values measured for each tissue sample on an array. At the same time, the sample size is typically far smaller than the data dimension. This situation necessitates dimensionality reduction through gene selection to avoid data over-fitting and improve generalization of discriminant analysis. In the context of cancer classification, a gene expression signature refers to a set of differentially expressed genes which, combined in a certain formalism, can discriminate one cancer from others. The objective of gene selection is to select those genes whose expressions define a signature for a particular cancer. Approaches to gene selection range from statistical analysis [4] and a Bayesian model [13] to Fisher's linear discriminant analysis [14] and support vector machines (SVMs) [15]. We notice, however, fundamental issues concerning reliability and diversity have not been addressed adequately. Both issues are critical as the process of gene selection may be sensitive to algorithmic parameters and data composition. To address these issues, we present a new method for multi-class cancer subtype classification that uses a cross-validation mechanism

*Corresponding author. P.O. Box 9706, Anaheim, CA 92812, USA. Fax: (1)-949-509 0230.
E-mail address: lifu@patcar.org (L.M. Fu).

0014-5793/04/$30.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(04)00175-9

to confer both reliability and diversity to selected genes. We show how to obtain reliable gene expression signatures for cancer subtype classification by conducting reliability analysis of selected genes. This method selected a smaller set of genes than previously reported techniques while maintaining the optimum level of predictive performance on the benchmark microarray data sets of SRBCTs and leukemia.

2. Materials and methods

2.1. Microarray gene expression data

The SRBCT data set includes 63 training samples and 25 test samples derived from both tumor biopsy and cell lines [2]. In consistency with other reports in the literature, we used the test set of 20 samples after five non-SRBCT samples were removed. The data set consists of four types of tumor in childhood, including Ewing's sarcoma (EWS), rhabdomyosarcoma (RMS), neuroblastoma (NB), and Burkitt lymphoma (BL). The data were obtained from cDNA microarrays. After initial screening, the data set in the public domain contains 2308 genes, and is available at http://research.nhgri.nih.gov/microarray/Supplement/.

The leukemia data consist of 72 tissue samples, each with 7129 gene expression values. The samples include 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML). The original data have been divided into a training set of 38 samples and a test set of 34 samples. The data were produced from Affymetrix gene chips. The data set is available at http://www-genome.wi.mit.edu/cancer/.

2.2. Gene ranking

Multivariate gene selection is best exemplified by a technique based on SVMs [15], which have been recognized as a powerful approach to classifier design [16,17]. The computational principle of a SVM is to find a particular hyperplane that offers the maximum possible separation between different classes of instances. The basic problem for training a SVM can be reformulated as: given a set of n training instances, each represented as $(\vec{x}_i, y_i)$ where $\vec{x}$ is the feature vector, y is the class label and $1 \le i \le n$, maximize

$$J = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n} y_i y_j \alpha_i \alpha_j (\vec{x}_i \cdot \vec{x}_j)$$

subject to

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \text{ and } \alpha_i \ge 0,\ 1 \le i \le n.$$

The optimal hyperplane that separates different classes of objects can be constructed from the solutions α_i's to this maximization problem. When the instances are not linearly separable, a soft-margin algorithm as an extension of the basic algorithm is available [18].

Gene ranks are determined by the SVM recursive feature elimination algorithm [15], in which the least important feature is identified and removed, remaining features are re-evaluated, and the process is repeated until no more features are available. Mathematically, genes are ranked by the absolute magnitude of the associated weight in the weight vector given by

$$\vec{w} = \sum_i \alpha_i y_i \vec{x}_i$$

The smaller the associated weight magnitude, the lower the rank of the gene. To re-rank the remaining genes, the SVM is trained again based on the data where the feature vector of each training instance is encoded only by the remaining genes.

2.3. Reliability analysis of gene selection

The innovative feature of our method is to conduct reliability analysis for arriving at the gene expression signature. The analysis assesses the repeatability of genes selected and determines the repeatability for gene selection using M-fold cross-validation. Cross-validation is a method for measuring the generalization performance of a machine learning or pattern recognition system, but the application of this technique to learning the pattern in the data is novel.

In the 10-fold cross-validation approach, the data set is divided into 10 disjoint subsets of about equal size. Genes are selected on the basis of nine of these subsets, and then the remaining subset is used to estimate the predictive error of the trained classifier using only the selected genes. This process is repeated 10 times, each time leaving one set out for testing and the others for training. The cross-validation error rate is given by the average of the 10 estimates of the error rate thus obtained.

In each cross-validation cycle, we conduct SVM-based gene ranking and selection operations. We select a minimal set of genes by collecting from the top rank one by one and picking the set associated with the minimum error rate with respect to the training data in each cross-validation cycle. There is no guarantee that the same subset of genes will be selected in each of the 10 cycles during the 10-fold cross-validation experiment. However, a vital gene tends to be selected consistently across cycles. The significance of a gene appears to be correlated with the repeatability of selection. We associate each selected gene with a repeatability value indicating how many times it is selected in the cross-validation experiment. The biological or clinical interpretation of 'repeatability' would depend on the objective and design of the microarray experiment. We consider the validity of a selected gene by its reliability. The reliability is measured by the repeatability that a gene is selected in the 10-fold cross-validation experiment. That is, the more often a gene is selected, the less likely chance is a factor.

In application of M-fold cross-validation to n samples, M can assume a value ranging from 2 to n. A small M is not sufficient to assess the repeatability of selected genes while a large M (e.g. $M = n$ in the leave-one-out experiment) is associated with a high degree of redundancy on data for training and low diversity of genes selected. As our experience shows, $M = 10$ is a good trade-off.

To select the final set of genes, we need to determine the repeatability threshold. A gene is in the final set if its repeatability reaches (i.e. no less than) the threshold. To this end, a second 10-fold cross-validation is performed. Then we choose the repeatability threshold associated with the minimal cross-validation error.

To extend the method from two-class to multi-class classification, we adopt the one-against-all-others strategy under which genes are selected for each class one at a time and then combined. For each class, all the other classes are grouped as a single class. In this way, a multi-class gene selection problem is converted into a series of two-class problems. The current program is available at http://www.cise.ufl.edu/~fu/NSF/cancer_classify_GES.html. The program was written in Matlab. An SVM Matlab toolbox as well as Mathlab is required for program use.

3. Results

The reference method with which we compared our method applied SVM-RFE to select genes from the whole training data without reliability assessment. The reference method [3] is a multi-class extension of the SVM-RFE method [15] used for two-class classification. The SVM-RFE method (two-class or multi-class) has not been applied to the SRBCT data. We implemented the computer algorithms of both methods for comparison. The SVM used in this study employed the linear kernel since we found that it yielded a better result than a non-linear kernel for the data under investigation, and it is also consistent with the literature [15]. All SVM parameters were set by default. The same experimental conditions were applied to both methods. For the SRBCT data, the Cy5/Cy3 fluorescence ratio data were $\log_{10}$-transformed; for leukemia data, the intensity of each gene on an array was divided by the mean intensity of all genes on that array in order to adjust variations between arrays.

3.1. SRBCT classification

On the SRBCT data, our method selected 32 genes (Table 1) from the microarray gene expression data of the 63 training samples. The SVM classifier trained on the 63 training sam-

Table 1
Genes selected by our method on the microarray dataset of SRBCTs

Image ID	Gene description	Tibshirani et al.	Khan et al.
21652	catenin (cadherin-associated protein), α1	●	●
298062	troponin T2, cardiac	●	●
383188	recoverin		●
755750	protein (NM23B) in non-metastatic cells 2		●
769716	neurofibromin 2	●	
878280	collapsin response mediator protein 1		●
377461	caveolin 1, caveolae protein	●	●
325182	cadherin 2, N-cadherin (neuronal)	●	●
1435862	MIC2 surface antigen (CD99)	●	●
42558	L-arginine:glycine amidinotransferase	●	●
812105	transmembrane protein	●	●
767183	hematopoietic cell-specific Lyn substrate 1		●
41591	meningioma 1	●	●
810057	cold shock domain protein A	●	
183337	major histocompatibility complex, class II, DM α	●	●
796258	sarcoglycan, α	●	●
1409509	troponin T1, skeletal, slow	●	●
788107	amphiphysin-like		●
143306	lymphocyte-specific protein 1		
866702	protein tyrosine phosphatase, non-receptor type 13	●	●
770394	Fc fragment of IgG, receptor, transporter, α	●	●
82225	secreted frizzled-related protein 1		●
52076	olfactomedin-related endoplasmic reticulum-localized protein	●	●
80109	major histocompatibility complex, class II, DQ α1	●	●
814260	follicular lymphoma variant translocation 1	●	●
784224	fibroblast growth factor receptor 4	●	●
204545	ESTs	●	●
244618	ESTs	●	●
295985	ESTs	●	●
308231	myh-1c		●
308163	ESTs	●	●
212542	cDNA DKFZp586J2118	●	●

Those genes also selected using the methods of Tibshirani et al. [12] and Khan et al. [2] are marked by the symbol ●.

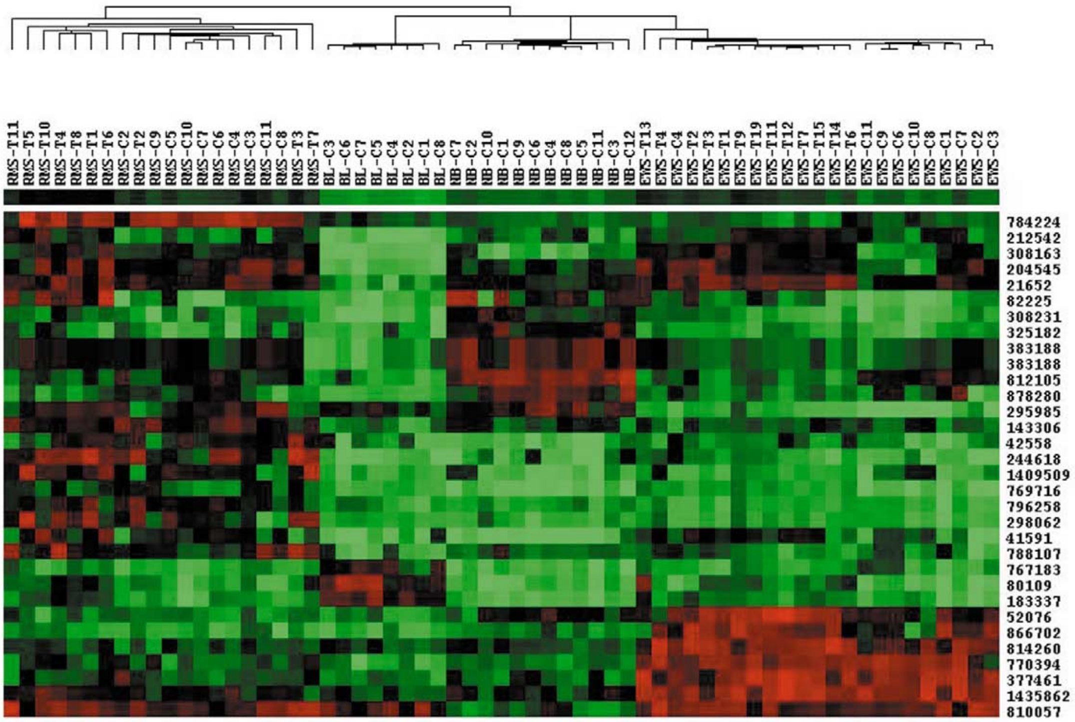

Fig. 1. The gene expression map of the 32 genes selected by our method for SRBCTs. The map was generated by Eisen's hierarchical clustering program called CLUSTER and viewed by the TREEVIEW program. Four sample clusters are visually recognized, corresponding exactly to the four predefined tumor classes (EWS, BL, NB, and RMS) with 100% accuracy.

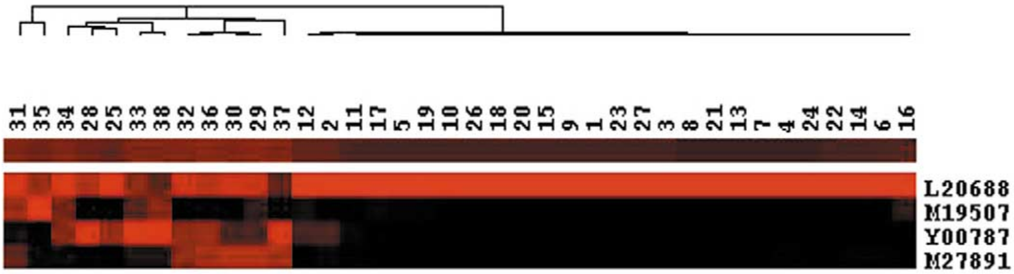

Fig. 2. The gene expression map of the four genes selected by our method for leukemia. AML and ALL samples are visually separable as two distinct clusters based on the gene expression profiles of the selected genes.

ples using the 32 selected genes was tested on the 20 different test samples. Both the training and test predictive accuracies were 100%. That is, the trained SVM classifier can accurately predict the tumor class using the 32 gene expression data for both seen and unseen samples. Since the classifier may tend to fit the training data, the generalization performance of the classifier is indicated by the test accuracy.

The reference method selected eight genes with 100% training accuracy but with only 90% test accuracy. It seemed that the reference method did not select enough genes even though the selected genes could correctly classify all the training samples – an example of data over-fitting, whereas the strategy of using multiple subsets of data in our method adequately dealt with this problem by taking into account both reliability and diversity in gene selection.

We examined the consensus of genes selected by our method and by two other best-known methods: the method of Khan et al. [2] based on artificial neural networks and the method of Tibshirani et al. [12] based on shrunken centroids, and we found that there was high consensus between our and their results. Out of the 32 genes selected by our method, 29 genes were also selected by Khan's method and 24 genes by Tibshirani's method. This is substantial evidence indicative of the validity and significance of our method.

Whether the selected genes served as meaningful markers for cancer classification was further confirmed by cluster analysis and visualization. In this regard, we applied a hierarchical clustering program developed by Eisen [19] to the gene expression data of the selected genes and then visualized the internal structure of the data (Fig. 1). The exact match between the gene expression clusters and the tumor classes attests to the soundness of our method. Later, we discuss the strength of our method over other methods.

3.2. *Leukemia classification*

On the leukemia data, our method selected four genes (Table 2) from the microarray gene expression data of 38 training samples. The SVM classifier trained on the 38 training samples using the selected genes was tested on the 34 different test samples. The training and test accuracies were 100% and 97.06%, respectively. In addition, the AML and ALL samples formed separate clusters in the gene expression map of the selected genes (Fig. 2).

The reference method also selected four genes and achieved the same level of test accuracy as our method. Recall, however, the reference method failed to give the optimum result on the SRBCT data. The original algorithm of SVM-RFE [15] selected eight or 16 genes on this data set without giving a criterion for breaking the tie. The method based on shrunken centroids [12] selected 21 genes on this data set. The best achievable unbiased test accuracy on the leukemia data seems to be 97.1%. Some studies reported 100% test accuracy because they combined training and test data for gene selection (bias selection) or because they normalized training and test data all together. A recent study indicated that the unbiased error estimate of the classifier using a small number of selected genes was virtually non-zero on the leukemia data set [20]. Taken together, the evidence showed that our method produced optimum results in terms of both predictive performance and the number of selected genes.

4. Discussion

In the context of cancer classification and discriminant analysis, gene selection methods are compared by the predictive performance and the number of genes selected. The goal of gene selection is to select a minimally required set of genes associated with optimum predictive performance. As part of validation, selected genes should be biologically important and offer insight into disease mechanisms. Some biologically significant genes are not selected if they are correlated with other even more significant genes. It should also be noted that the valid number of selected genes is constrained by the available sample size.

The two best-known techniques tested on the SRBCT microarray data set are based on artificial neural networks [2] and statistical shrunken centroids [12], which selected 96 and 43 genes, respectively, both with 100% accuracy on 63 training and 20 test samples. In comparison, our method selected a smaller set of 32 genes with the same level of perfor-

Table 2
Genes selected by our method on the leukemia microarray dataset

Accession number	Gene description	Golub et al.	SVM-RFE
M27891	CST3 cystatin C	●	●
Y00787	interleukin-8 precursor	●	●
M19507	MPO myeloperoxidase		●
L20688	Ly-GDI		

Those genes also selected using the methods of Golub et al. [4] and SVM-RFE (the reference algorithm) are marked by the symbol ●.

mance (100% accuracy on both training and test data), allowing for a more cost-effective classifier. The above three techniques including ours selected genes and trained the classifier using the 63 training samples, and the performance of the classifier was tested on the 20 test samples. In contrast, another technique based on statistical within-class variation mixed the 63 training and 20 test samples and performed repeated five-fold cross-validation, resulting in a non-zero cross-validation error and a smaller set of 21 selected genes; selected genes were further evaluated using leave-one-out on the same data set from which the genes were selected [21].

Genes selected by our method for a particular type of cancer/tumor against other types are generally consistent with its tissue of origin. For example, genes selected for NB are characteristic for nerve cells, such as recoverin, neurofibromin 2, neuronal N-cadherin, and meningioma 1; genes selected for RMS are characteristic for muscle cells, such as cardiac troponin T2, α-sarcoglycan, and slow skeletal troponin T1; genes selected for BL are characteristic for lymphocytes or blood cells, such as hematopoietic cell-specific Lyn substrate 1, major histocompatibility complex class II, DM α, and major histocompatibility complex class II, DQ α1. Some genes discovered by means of microarray analysis have been reported in the biological literature, e.g. over-expression of MIC2 in EWS [22]. Some genes are over-expressed in a certain type of tumor but lack specificity. For instance, FGFR4 (fibroblast growth factor receptor 4) was noted to be highly expressed only in RMS and not in normal muscle, but it is also expressed in some other cancers and normal tissues [2]. A gene that is under-expressed in a particular type of tumor compared with other types can also be selected as a diagnostic marker. For instance, cold shock domain protein A selected for NB was under-expressed in this tumor, consistent with the fact that this gene is expressed in B cells and skeletal muscle but not in the brain [12].

The confidence in the validity of a selected gene is increased if it is selected by multiple techniques. This justifies the approach in which gene selection is based on the consensus of multiple techniques. On the SRBCT data set, the consensus of artificial neural networks [2], shrunken centroids [12], and our method selected 22 genes with 100% accuracy on both training and test data. Thus, it appears to be a good idea in this case, though it can be argued that the consensus approach may impose an overly stringent criterion and end up selecting a less than optimum number of genes.

We emphasize the importance of holding back some data to improve generalization and diversity of the learning outcome. The distinctive feature of our method is that gene selection is determined by both ranking and reliability analyses. Reliability analysis is conducted using M-fold cross-validation. Some gene selection methods [15,21] use cross-validation to determine the number of selected genes by the minimum cross-validation error but not by the optimum repeatability as in our method. Thus, reliability analysis comprising repeatability measurement and optimum repeatability determination defines the novelty of our method, which has enabled a more accurate and cost-effective cancer classifier to be constructed, compared with the reference method and other methods. The discovered novel genes that characterize a cancer type may suggest new molecular targets for drug discovery in addition to their diagnostic and prognostic significance in cancer research and management.

Acknowledgements: This work is supported by the National Science Foundation under Grant IIS-0221954.

References

[1] Sorensen, J.B., Hirsch, F.R., Gazdar, A. and Olsen, J.E. (1993) Cancer 71, 2971–2976.
[2] Khan, J. et al. (2001) Nat. Med. 7, 673–679.
[3] Ramaswamy, S. et al. (2001) Proc. Natl. Acad. Sci. USA 98, 15149–15154.
[4] Golub, T.R. et al. (1999) Science 286, 531–537.
[5] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Proc. Natl. Acad. Sci. USA 96, 6745–6750.
[6] Perou, C.M. et al. (1999) Proc. Natl. Acad. Sci. USA 96, 9212–9217.
[7] Brooks, J.D. (2002) Curr. Opin. Urol. 12, 395–399.
[8] Bremnes, R.M. et al. (2002) J. Clin. Oncol. 20, 2417–2428.
[9] Clarke, P.A., te Poele, R., Wooster, R. and Workman, P. (2001) Biochem. Pharmacol. 62, 1311–1336.
[10] Cooper, C.S. (2001) Breast Cancer Res. 3, 158–175.
[11] Habeck, M. (2001) Lancet Oncol. 2, 5.
[12] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Proc. Natl. Acad. Sci. USA 99, 6567–6572.
[13] Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M. and Mallick, B.K. (2003) Bioinformatics 19, 90–97.
[14] Xiong, M., Li, W., Zhao, J., Jin, L. and Boerwinkle, E. (2001) Mol. Genet. Metab. 73, 239–247.
[15] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Machine Learning 46, 389–422.
[16] Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares Jr., M. and Haussler, D. (2000) Proc. Natl. Acad. Sci. USA 97, 262–267.
[17] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) Bioinformatics 16, 906–914.
[18] Cortes, C. and Vapnik, V. (1995) Machine Learning 20, 273–297.
[19] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14868.
[20] Ambroise, C. and McLachlan, G.J. (2002) Proc. Natl. Acad. Sci. USA 99, 6562–6566.
[21] Cho, J.H., Lee, D., Park, J.H. and Lee, I.B. (2003) FEBS Lett. 551, 3–7.
[22] Kovar, H., Dworzak, M., Strehl, S., Schnell, E., Ambros, I.M., Ambros, P.F. and Gadner, H. (1990) Oncogene 5, 1067–1070.