PROBABILISTIC ARTIFICIAL NEURAL NETWORKS FOR TISSUE CLASSIFICATION USING GENE EXPRESSION PROFILES

C. Papaloukas*,**, T. Nikolaidi*, Y. Goletsis*,*** and D.I. Fotiadis*

* Unit of Medical Technology and Intelligent Information Systems, Department of Computer Science, University of Ioannina, Ioannina, GR45110, Greece
** Department of Biological Applications and Technology, University of Ioannina, Ioannina, GR45110, Greece
*** Department of Economics, University of Ioannina, Ioannina, GR45110, Greece

fotiadis@cs.uoi.gr

Abstract: Gene expression profiles hold valuable information providing understanding for many cellular processes. The utilization of microarrays offers the possibility of monitoring gene expression for tens of thousands of genes in parallel. In this work, microarray data were used under the framework of a radial basis functions artificial neural network for the classification of two common types of cancer: acute myeloid leukaemia and acute lymphocytic leukaemia. The classification procedure is implemented according to a four-stage schema: neighbourhood analysis, selection of the informative genes, boosting and, finally, classification of the given tissue. To evaluate the proposed method we used previously reported data consisting of 72 leukaemia samples. Our method seems to outperform other similar systems, since after being trained with only 12 samples, instead of 38 like all the other methods do, it manages to correctly classify the 59 out of the 60 samples, yielding a sensitivity of 100% and a specificity of 94%. When ROC analysis is employed to evaluate the diagnostic approach, the area under the curve had a value of 0.99, approximately.

Introduction

Each tissue is characterized at a given time by a unique pattern of gene expression, which is called "expression profile" or "molecular signature". These unique expression profiles of different tissues or of the same tissue under different experimental conditions hold valuable information which provides understanding and insight into many cellular processes. Microarray analysis provides the possibility of monitoring gene expression for tens of thousands of genes in parallel and is expected to significantly contribute to the development of efficient cancer diagnosis approaches.

The main steps of this analysis include data normalization and filtering in order to establish the differential gene expression [1], dimensionality reduction [2] and finally, pattern recognition that will assign biological meaning to the expression profiles. Pattern identification demands further analysis of the microarray data, which includes gene identification [3], gene regulatory network modelling [4], clustering and classification. The last two are mainly implemented by machine learning approaches.

Clustering techniques applied in microarray analysis include hierarchical clustering algorithms [5], self-organizing maps [6], and graph theoretical approaches [7,8]. For the purpose of classification, the techniques applied were linear discriminant analysis [9-12], k nearest neighbour classifiers [13], support vector machines [14,15], self-organizing maps [16], boosting algorithms [7,16], decision trees [17,18], multilayer perceptrons [18] and others [19-23].

In the present work microarray analysis is combined with the use of Artificial Neural Networks (ANNs) for the classification of cancer. More specifically, probabilistic neural networks (PNNs) are employed to classify two types of cancer: acute myeloid leukaemia (AML) and acute lymphocytic leukaemia (ALL). For training and testing the method we used the dataset derived from [16], as well as, the same set of informative genes selected at this study. We also applied a simple boosting algorithm. Eventually, the developed diagnostic approach classified the two types of cancer indicating high accuracy of the method.

Materials and Methods

Microarray Analysis

Microarrays allow monitoring of gene expression for thousands of genes, by conducting massively parallel hybridization experiments under particular experimental conditions and environments. Thus, they produce huge amounts of valuable data and assist in the identification of novel genes or associate genes within complex gene pathways [24]. The gene expression patterns derived from microarray hybridization experiments provide snapshots of the state of a living cell, which determine its biological behaviour (Figure 1). If we consider the fact that a human cell contains approximately 3 billion base pairs, which encode about 50,000 to 100,000 genes and only a fraction of these genes are expressed in any given tissue, instead of treating gene expression pattern from a given microarray experiment as a single data entity, we can examine one gene at a time across a

ISSN: 1727-1983

biological process or a collection of biological samples (the gene expression profile).

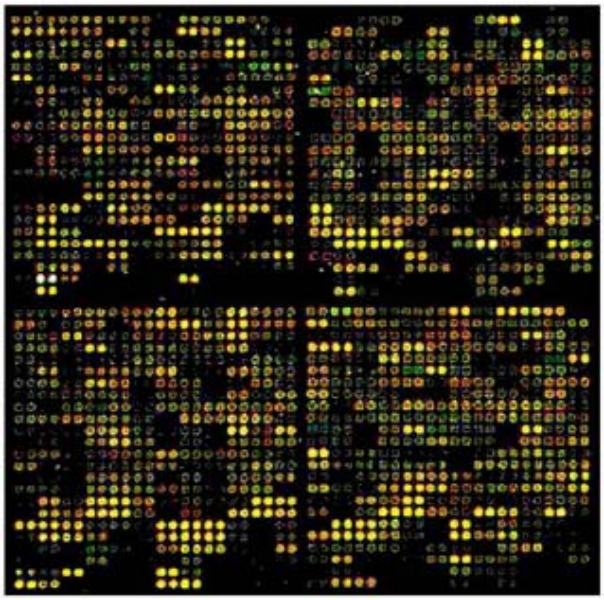

Figure 1: A microarray
(Source: http:// plantgenomics.biology.yale.edu/).

The analogy between biological sequences and microarray expression data is not directly evident. While a protein sequence can be represented as a string of letters drawn from an alphabet of twenty amino acids, microarray expression measurements are usually represented as a fixed-length vector of continuous values [25]. For cDNA arrays, each value is the logarithm of the ratios of the estimated abundances of mRNA in two tissue samples. A microarray experiment produces about 10,000 such ratios, each generally corresponding to a particular gene, so that gene expression patterns can be compared by determining the ratio of fluorescent intensities of the two dyes after hybridisation with the probes. This ratio, see Eq. (1), is the log ratio between the two intensities and provides the gene expression value [5,26]:

$$Gene\_Expression = \log_2 \frac{Intensity(Cy5)}{Intensity(Cy3)}. \qquad (1)$$

Classification Problem

Leukaemia is a form of cancer that affects bone marrow and the production of white blood cells. The disease exists in several forms, but is always malignant. AML occurs when immature blood cells, the myeloblasts, fail to follow the differentiation, the process by which the healthy myeloblasts mature into white blood cells, platelets, and red blood cells. The failure of differentiation results in the accumulation of the myeloblasts in the bloodstream, where they eventually count more than healthy blood cells. From the bloodstream, the cancer can spread into the liver, the spleen, or any other organ in the body. AML is associated with chromosomal translocations, where genetic information on one chromosome switches places with information on another chromosome. Particularly, AML is associated with the t(8;21)(q22;q22) translocation, that occurs in 15% of patients with AML [16].

ALL occurs when the immature lymphocytes, or blasts, fail to develop into mature white blood cells, and accumulate in the blood stream and bone marrow. Those abnormal lymphocytes that cannot mature are often larger than the normal. The rapid reproduction of the abnormal lymphocytes results in high white blood cell counts, and low levels of red blood cells. ALL is also associated with certain chromosomal translocations: the t(12;21) (p13;q22) translocation that occurs in 25% of patients with ALL [16].

The fact that chemotherapy regimens for AML relies on a backbone of daunorubicin and cytarabine, while for ALL generally contains corticosteroids, vincristine, methotrexate, and L-asparaginase, indicates that a critical issue for the successful treatment is to distinguish ALL from AML.

Classification Procedure

The classification of leukaemia based on PNNs was implemented according to the schematic diagram shown in Fig. 2. First, neighbourhood analysis of the microarray data that include expression levels of 6,817 genes is performed; this is followed by the selection of the informative genes (50 more representative ones); then the selection of the training and test sets according to a boosting algorithm takes place and, finally, the probabilistic neural network is applied for the the classification of the given tissue.

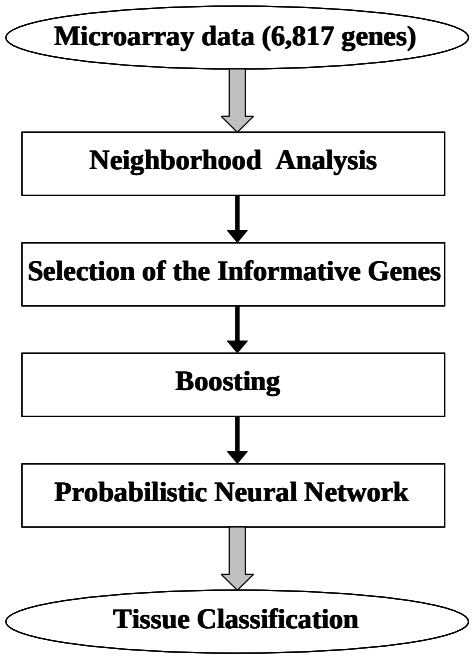

Figure 2: The four stages of the tissue classification procedure.

ISSN: 1727-1983

Microarray Data

In our work we used the data generated by a similar study in [16]. Their initial leukaemia training data set consisted of 27 ALL, and 11 AML, i.e. 38 bone marrow samples obtained from acute leukaemia patients at the time of diagnosis before chemotherapy. Samples were randomly selected from the leukaemia cell bank based on availability. The 27 ALL samples were obtained from ALL childhood patients treated on Dana-Farber Cancer Institute (DFCI) and the 11 adult AML samples derived from the Cancer and Leukaemia Group B (CALGB) protocols between 1980 and 1999. All the samples were selected regardless to cytogenetics, immunophenotype, or other molecular features. The RNA derived from bone marrow mononuclear cells was hybridised to high-density oligonucleotide microarrays, that included probes for 6,817 human genes and a quantitative expression level was obtained for each gene. This dataset is available at: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

Neighbourhood Analysis

In the first stage of the classification procedure the identification of the genes, whose expression pattern is strongly correlated with the class distinction of the leukaemia samples to AML and ALL, is realised [16]. The 6,817 genes are sorted by their degree of correlation and the "neighborhood analysis" method is employed to establish whether the observed correlations are stronger than would be expected by chance.

Each gene is represented by an expression vector $v(g)=(e_1, e_2, ..., e_n)$, where e_i denotes the expression level of gene g in the i-th sample in the initial set S of samples. An ideal expression pattern $c=(c_1, c_2, ..., c_n)$, represents the class distinction where $c_i=1$ or 0 according to whether the i-th sample belongs to class 1 or class 2. $P(g,c)$ was the measure of correlation used, in a way that large values of $|P(g,c)|$ indicate a strong correlation between the gene expression and the class distinction, while the sign of $P(g,c)$ being positive or negative corresponds to g being more highly expressed in class 1 or class 2. The sets of genes such that $P(g,c)=r$ and $P(g,c)=2r$ are defined as neighbourhoods $N_1(c,r)$ and $N_2(c,r)$ of radius r around class 1 and class 2, respectively. If the number of genes within the neighbourhoods is large, many genes will have expression patterns closely correlated with the class vector.

Selection of the informative genes

When the above method of neighbourhood analysis was applied to the 38 acute leukaemia samples, roughly 1,100 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance. From the 1,100 genes the 50 were chosen arbitrarily, and were called the "informative genes" [16]. Those could be the genes we must use for the classification of our samples, as well.

Boosting

One approach to boosting is the heuristic selection of a small number of prototypes that performs equally well with the complete dataset [27]. Bearing this in mind, we selected the samples, as well as, their number, following a try-and-error approach. Specifically, in our study, in contrary with [16], we did not split the 72 samples of the overall dataset into 38 training and 34 test samples. We trained the system with different number of samples all including the 50 genes selected from the work in [16] with neighbourhood analysis.

Probabilistic Neural Networks

An artificial neural network simulates multiple layers of simple processing elements, the neurons. It is usually composed of a great number of interconnected artificial neurons that are simplified models of their biological counterparts. The neurons are linked to many of its neighbours and the coefficients of connectivity represent the strengths of these connections. The components of neural networks are modelled according to the structure of the brain and in general neural networks have a strong similarity to the biological brain.

We used radial basis functions ANNs for the classification of the two types of cancer, AML and ALL. In a radial basis network the input to the transfer function (activation of a neuron) is the vector distance between its weight vector w and the input vector p, multiplied by the bias b. The radial basis function has a maximum value equal to 1 when its input is 0. As the distance between w and p decreases, the output increases. Thus, a radial basis neuron acts as a detector that produces 1 whenever the input p is identical to its weight vector w. As for the bias b, it allows the sensitivity of the radial basis neuron to be adjusted [28].

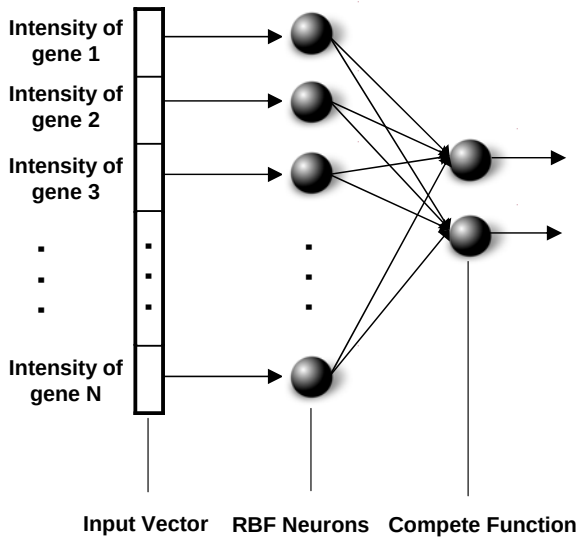

Figure 3: Architecture of the Probabilistic Neural Network.

ISSN: 1727-1983

Our work is based on probabilistic neural networks which belong to a subclass of radial basis functions ANNs and can be used for classification problems. When an input is given, the first layer computes distances from the input vector to the training input vectors, and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce a vector of probabilities. Finally, a "compete" transfer function on the output of the second layer picks the maximum of these probabilities, and produces 1 for that class and a 0 for the other classes [28]. The architecture for the ANN is shown in Fig. 3. It should be noted that the input to the PNN are the intensity values of each gene for each sample derived from the microarray experiment image and the output is the classification of the samples into the two cancer classes, AML and ALL.

Results

The sensitivity (*Se*) and specificity (*Sp*) measures were employed to assess the performance of the proposed classifier. We started training the system with 2 samples, and following a resampling procedure (i.e. the boosting algorithm) we continuously increased the number of samples by adding a pair of samples at a time, one AML and one ALL. The pairing of the samples was adopted in order to reduce during training the bias in one of the two classes. We observed that after 12 samples, there is no change in the results (*Sp* drops since the number of test samples decreases), fact that proves that we can train the system quite well with only 12 samples. From the 72 samples the remaining 60 were used for testing, and as seen in Table 1, only one out of the 60 was misclassified.

Table 1: Classification results for different number of training samples.

Training samples	TP	TN	FP	FN	*Se* (%)	*Sp* (%)
6	44	18	4	0	100	81,82
8	43	18	2	1	97,73	90
10	42	17	1	2	95,45	94,44
12	43	16	1	0	100	94,12
14	43	14	1	0	100	93,33
16	42	13	1	0	100	92,86
18	41	12	1	0	100	92,3
20	39	12	1	0	100	92,3
30	29	12	1	0	100	92,3
40	22	9	1	0	100	90
50	15	6	1	0	100	85,71

The *Se* and *Sp* values (those produced from the 12-sample training set) were used for the estimation of the ROC curve shown in Figure 4. The estimated area under curve was very high (AUC = 0.9898), fact that proves the high performance of the proposed classification method.

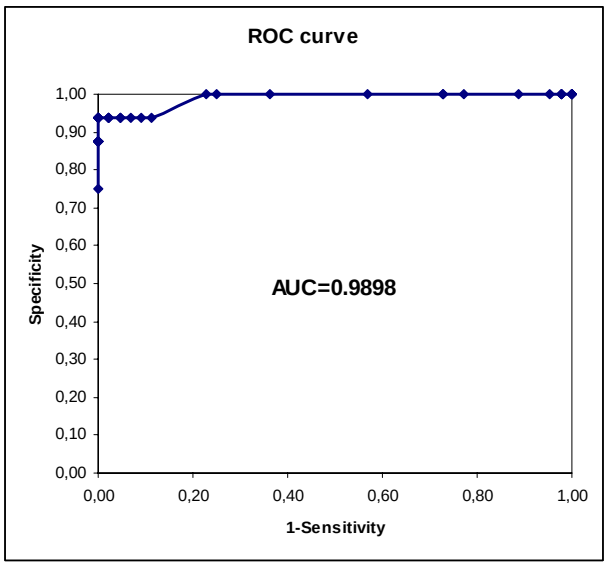

Figure 4: Receiver Operating Characteristic (ROC) curve.

Discussion

A new method was developed based on ANNs for tissue classification purposes using gene expression data derived from a microarray experiment. More specifically, we employed PNNs which are a type of radial basis functions ANNs. The proposed method aims at classifying two types of cancer: AML and ALL, and proved to be very reliable, since it misclassified only one sample from the test set.

As already has been discussed in [16] the original 72 samples were split in a 38-training set and a 34-test set. Their neighbourhood analysis correctly identified 29 samples with strong predictions, and misclassified 2 of the other 5 samples that had weak predictions. In [10] linear discriminant analysis was employed for the same classification problem. Following the approach of leave-one-out-cross-validation, they selected 8 genes that yielded no misclassifications. Utilizing a similar methodology in [11] and specifically Fisher linear discriminant analysis combined with forward stepwise feature selection again all the samples were correctly classified. It should be noted that cross validation was also employed in this work. Likewise, step-wise cross-validated discriminant analysis was used in [12] which was tested on an independent dataset and managed to correctly classify 32 of the 34 samples. On the other hand, support vector machines [14] produced classification results that ranged from 30-32 accurate predictions on the same 34-sample dataset. Their results depend on the number of genes they used that varied from 25 - 1000. In [19] the concept of emerging patterns was employed for the classification of the AML-ALL dataset. They used only one gene (zyxin) for classification and managed to correctly classify 31 of the 34 test samples. Between group analysis of microarray data was another technique proposed in [20], which identified 25 discriminating genes. Those were

ISSN: 1727-1983

Table 2: Classification performance of various methods on the same dataset with leukaemia samples.

Method	Number of informative genes	Accuracy	(%)
Linear discriminant analysis [10]	8	72/72	100
Fisher linear discriminant analysis & Forward stepwise selection [11]	2	72/72	100
Step-wise cross-validated discriminant analysis [12]	2	32/34	94
Support vector machines [14]	25-1000	30-32/34	88-94
Weighted voting [16]	50	29/34	85
Emerging patterns [19]	1	31/34	91
Between group analysis [20]	25	30/34	88
Independently consistent expression discriminator [22]	16	34/34	100
Bayesian model averaging [23]	20	32/34	94
Probabilistic neural network	**50**	**59/60**	**98**

used for the tissue classification, yielding a correct classification of 30 to 33 (depending on the filtering of the genes) out of the 34 samples. In [22] a different approach was applied, called independent consistent expression discriminator, and used 38 samples to develop a weighting-voting system that selected 16 genes, instead of the 50 that were selected in [16]. Their method correctly identified all the 34 test samples. Finally, Bayesian model averaging [23] after selecting a set of 20 genes classified correctly 32 of the 34 samples.

Our method seems to outperform the above systems (Table 2), since it manages to correctly classify 59 of the 60 samples, after being trained with only 12 samples, instead of 38 like all the other methods do. The one misclassification is false positive, which in classification problems between normal and cancerous tissues false positives are considered to be tolerable. Thus, the proposed method will perform even better in that type of discrimination problems.

During the training and testing of the system we observed that there are some samples in the dataset that are more discriminative than the rest and have to be presented first for training in the ANN classifier. More specifically, the samples in the following order {1, 72, 38, 68, 28, 50, 37, 64, 32, 66, 35, 58} gave the best results, while a different order would decrease the classification performance. This has to do with the architecture of the PNN and the small number of samples in the employed dataset. A preprocessing stage (for example the application of a k-nearest neighbour algorithm) would automate the boosting procedure yielding a fully automated classification system.

Finally, in the latest studies of cancer classification, the number of genes that are used as the set of the informative genes is gradually reducing for the same dataset. In particular, there are studies that attempt to discriminate the two types of cancer by using 25, 16, 8, 5, or even only one gene. Given the fact that their performance is comparable with other more reliable methods [14,16] and following the principle of Okam's razor we should also apply our method on samples with smaller number of genes.

Conclusions

We presented a new method for leukaemia classification based on PNNs and microarray data. The proposed method proved to be very reliable since it misclassified only one sample out of 60. Furthermore, it can be trained with a minimal set of samples, in the leukaemia example with only 12, which is very desirable in classification problems where the dataset is limited as is the case with the microarray data. Further testing of the proposed PNN diagnostic approach with other types of cancer will fully reveal the overall efficacy of the method.

References

[1] QUACKENBUSH J. (2001): 'Computational genetics computational analysis of microarray data', *Nat. Rev. Genet.*, **2**, pp. 418-27

[2] ALTER O., BROWN P.O., BOTSTEIN D. (2000): 'Singular value decomposition for genome-wide expression data processing and modeling', Proc. Natl. Acad. Sci. U.S.A., pp. 10101-6

[3] 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**, pp. 5-14

[4] FRIEDMAN N., LINIAL M., NACHMAN I. and PEER D. (2000): 'Using Bayesian networks to analyze expression data', *J. Comput. Biol.*, **7**, pp. 601-20

[5] EISEN M.B., SPELLMAN P.T., BROWN P.O. and BOTSTEIN D. (1998): 'Cluster analysis and display of genome-wide expression patterns', Proc. Natl. Acad. Sci. U S A, pp. 14863-8

[6] TORONEN P., KOLEHMAINEN M., WONG G. and CASTREN E. (1999): 'Analysis of gene expression data using self-organizing maps', *FEBS L.*, **451**, pp. 142-6

ISSN: 1727-1983

[7] BEN-DOR A., BRUHN L., FRIEDMAN N., NACHMAN I., SCHUMMER M. and YAKHINI N. (2000): 'Tissue classification with gene expression profiles', *J. Comput. Biol.*, **7**, pp. 559-84

[8] HARTUV E., SCHMITT A., LANGE J., MEIER-EWERT S., LEHRACH H. and SHAMIR R. (2000): 'An algorithm for clustering cDNA fingerprints', *Genomics*, **66**, pp. 249-56

[9] 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

[10] CHO J.H., LEE D., PARK J.H., KIM K. and LEE I.B. (2002): 'Optimal Approach for Classification of Acute Leukemia Subtypes Based on Gene Expression Data', *Biotechnol. Prog.*, **18**, pp. 847-54

[11] XIONG M., LI W., ZHAO J., JIN L. and BOERWINKLE E. (2001): 'Feature (gene) selection in gene expression-based tumor classification', *Mol. Genet. Metab.*, **73**, pp. 239-47

[12] SOUKUP M., LEE J.K. (2004): 'Developing optimal prediction models for cancer classification using gene expression data', *J. Bioinform. Comput. Biol.*, **1**, pp. 681-94

[13] SBONER A., ECCHER C., BLANZIERI E., BAUER P., CRISTOFOLINI M., ZUMIANI G. and FORTI S. (2003): 'A multiple classifier system for early melanoma diagnosis', *Artif Intell Med.*, **27**, pp. 29-44

[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**, pp. 906-14

[15] 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 Natl. Acad. of Sci. U.S.A., pp. 262-267

[16] 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**, pp. 531-537

[17] 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**, pp. 673-79

[18] 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**, pp. 3493-7

[19] LI J. and WONG L. (2002): 'Identifying good diagnostic gene groups from expression profiles using the concept of emerging patterns', *Bioinformatics*, **18**, pp. 725-34

[20] CULHANE C.A., PERRIERE G., CONSIDINE C.E., COTTER G.T. and HIGGINS G.D. (2002): 'Between-group analysis of microarray data', *Bioinformatics*, **18**, pp. 1600-8

[21] LI Y., CAMPBELL C., and TIPPING M. (2002): 'Bayesian automatic relevance determination algorithms for classifying gene expression data', *Bioinformatics*, **18**, pp. 1332-9

[22] BIJLANI R., CHENG Y., PEARCE A.D., BROOKS I.A. and OGIHARA M. (2003): 'Prediction of biologically significant components from microarray data: Independently Consistent Expression Discriminator (ICED)', *Bioinformatics*, **19**, pp. 62-70

[23] YEUNG K.Y., BUMGARNER R.E. and RAFTERY A.E. (2005): 'Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data', *Bioinformatics*, **21**, pp. 2394-402

[24] HARRINGTON C.A., ROSENOW C. and RETIEF J. (2000): 'Monitoring gene expression using DNA microarrays', *Curr. Opin. Microbiol.*, **3**, pp. 285-91

[25] PAVLIDIS P., TANG C. and NOBLE W.S. (2001): 'Classification of genes using probabilistic models of microarray expression profiles', Proc. of BIOKDD: Workshop on Data Mining in Bioinformatics

[26] DERISI J.L., IYER V.R. and BROWN P.O. (1997): 'Exploring the metabolic and genetic control of gene expression on a genomic scale', *Science*, **278**, pp. 680-6

[27] FREUND Y. and SCHAPIRE R.E. (1996): 'Experiments with a New Boosting Algorithm', Proc. of 13th Intern. Conf. on Machine Learning

[28] DEMUTH H. and BEALE M. (2000): 'Neural Network Toolbox User's Guide', The Mathworks, Inc

ISSN: 1727-1983