Human Molecular Genetics, 2003, Vol. 12, No. 8 **823–836**
DOI: 10.1093/hmg/ddg093

Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification

Chiara Romualdi, Stefano Campanaro, Davide Campagna, Barbara Celegato, Nicola Cannata, Stefano Toppo, Giorgio Valle and Gerolamo Lanfranchi*

CRIBI Biotechnology Centre and Dipartimento di Biologia, Università degli Studi di Padova,
Via Ugo Bassi 58/B, 35121 Padova, Italy

Received November 11, 2002; Revised and Accepted February 7, 2003

Large-scale parallel measurements of the expression of many thousands genes are now available with high-density array made with collections of cDNA fragments, or oligonucleotide corresponding to different transcripts. These technologies have been applied to cancer investigations since the availability of such a large number of markers makes DNA array a powerful diagnostic tool for tumour and patient classification. Over the last two years, a series of computational tools have been developed for the analysis of different aspects of gene profiling. Our work tries to compare a series of supervised statistical techniques on the basis of their ability to correctly classify different types of tumours. A simulation approach was initially used to control the huge source of variation among and between patients, and to evaluate the ability of algorithms to classify tumours in relation to different types of experimental variables. Different techniques for reduction of data dimension were then added to the discriminant analysis and compared according to their ability to capture the main genetic information. The simulation results have been tested by applying the selected classification algorithms to two experimental microarray datasets of human cancers, and by measuring the correspondent rates of misclassification. Our analyses identify in these datasets a series of genes principally involved in tumour characterization. The functional role of these discriminant transcripts is discussed.

INTRODUCTION

A variety of molecular, clinical and morphological parameters are currently used to distinguish and classify human neoplasia and affected patients into discrete classes for a more accurate diagnosis and treatment. These parameters are derived from a series of established histological, immunological and biochemical techniques for tumour analysis. Despite the advancement of these technologies and the enhancement of their level of detection, the sensitivity and degree of prediction of cancer evolution need further improvements. The recent experimental technologies based on DNA arrays offer the possibility to study the expression levels of thousands of genes simultaneously (1). With this approach it will be possible to discover hundreds of novel molecular markers that can lead to a finer definition of tumour diversity (2). A careful analysis of the gene expression patterns through different diagnostic situations may help to classify patients into the correct tumour category and/or to detect new tumour stratifications. This is extremely important because neoplasias that are traditionally classified into homogeneous groups, often evolve into diverse clinical outcomes, and respond differently to pharmacological treatments. Global gene expression studies can lead to the definition of the molecular diversity underlying such phenomena, discovering marker genes whose under- or over-expression could be connected to novel traits in apparently uniform classes of tumours.

Several studies have already shown how DNA arrays can be effectively applied to distinguish different types of tumours according to their gene expression patterns (3–6). In spite of these recent progresses, many uncertainties still remain in cancer diagnosis. There are more than a hundred types of tumour (7) and, potentially, each type might be further divided into subtypes. In this context, the statistical methodologies used for classification become of crucial importance to recognize differences in the molecular structure between pathological

*To whom correspondence should be addressed. Tel: +39 0498276221; Fax: +39 0498276259; Email: lanfra@cribi.unipd.it samples and controls, and among series of cases belonging to the same diagnosed cancer type.

Different statistical methodologies may be used for the analysis of DNA array data for various aspects of cancer classification. The identification of new tumour types has been improved using unsupervised statistical techniques, such as cluster analysis (8,9). Prediction and categorization of cancer types into known classes have become possible thank to supervised methodologies such as discriminant analysis (3,10–14). However, supervised techniques should be considered as the final step after an accurate statistical analysis for tumour class definition. The accuracy of prediction algorithms is in fact highly influenced by the precision with which different tumour classes have been recognized and categorized. Finally, variable selection, also known as gene scoring, can be used either to identify putative genes whose expression pattern is highly related to specific tumour types, or to reduce the dimensionality of the data. In fact, it is well known that the performance of a given decision rule does not keep improving as the dimension of the features (genes) increases (15). Better discriminant results may be obtained if a feature vector of substantially lower dimension is used. Recently, many gene-scoring methodologies have been proposed (16–19). They may be divided into methods of single gene selection and methods of factors construction based on the generation of linear combinations of the original variables.

Clearly, not all the different types of discriminant analyses perform equally. A comparison of some statistical discriminant techniques based on the use of published expression datasets has been proposed (20). This study revealed that the traditional linear classifiers, like the diagonalized linear discriminant and the nearest-neighbour methodologies, perform better than more sophisticated processes such as the aggregated classification tree. Moreover, the actual performance of a given statistical approach for DNA array data analysis can be strongly affected by the structural variables of the experimental system approached (e.g. number of patients, gene expression patterns, or number of different cancer classes) and by the techniques used for features reduction. The influence of experimental variables and of dimension reduction techniques on classification results can be effectively controlled through a simulation approach.

The intention of our work was the assessment of the performance variability of various discriminant analyses in relation to: (i) different experimental variables to which the DNA array technology is applied for cancer classification; and (ii) the influence of different systems for decrease of data dimension. We have challenged six different classification algorithms that have been applied recently for the interpretation of cancer cDNA array data, and four different techniques for reduction of data dimension. We have tried to determine which statistical algorithms give the best performance in various experimental situations. Our simulation approach is based on the generation of several expression matrices that are representative of many different experimental conditions. Discriminant algorithms were applied to each of these simulated situations, and their behaviour was compared through error rates obtained with 10-fold cross validation. We found that partial least square (PLS) analysis highly increases the classification performance of all methodologies. In particular, with the aid of the PLS reduction, the neural network (NNET) is the algorithm that performs better when dealing with a small number of patients per tumour class, while the diagonalized linear discriminant analysis (LDA) performs better with a large number of tumour classes. We show that all the methodologies have comparable performances when the number of patients per tumour is greater than 50, the number of tumours is lower than four and the number of discriminating genes is larger than 40.

Our analyses have been applied, and the results can be referred to expression data obtained with microarray platforms made with the ‘deposition’ technology, which is the most common platform used in the scientific literature. This means that amplified cDNA inserts or synthetic oligonucleotides of different lengths corresponding to specific transcripts are arrayed at high density on glass slides. The Affimetrix DNA silicon chip (21), the second most used platform for expression studies, relies on the synthesis of a parallel series of matched and mismatched short oligonucleotides for each transcript directly into the silicon chip. The gene expression values obtained with Affimetrix platform are usually calculated as the average difference of the matched and mismatched intensities. Using average differences rather than ratio intensities, it is possible to obtain negative expression values. Since our simulation approach is based on random generation from a gamma distribution (a probability density function suitable only for positive variables), at the moment we have decided to focus our comparison to ‘deposition’ array data. We are, however, developing a similar approach to be applied particularly to expression data obtained with Affimetrix platforms.

The results obtained with the six selected classification algorithms using the simulation approach were tested using two published cDNA microarray datasets: the round blue cell tumours dataset (3) and the National Cancer Institute dataset (4,5). Exploiting the PLS factors obtained with dimension reduction techniques (22,23), we also obtained from these expression datasets lists of genes (associated with their functions) that are putatively responsible of tumour discrimination.

Supplementary Material about simulation approach and statistical results is available at http://muscle.cribi.unipd.it/microarrays/simulation/.

RESULTS

We have compared the performances of six supervised learning machines on simulated and experimental datasets of gene expression profiling. The classification algorithms were chosen from those recently proposed and applied in scientific literature for the interpretation of cancer gene expression profiling data obtained with DNA array. They are: the diagonalized linear discriminant analysis (LDA); the neural networks (NNET); the recursive partitioning (RPART); the support vector machine (SVM); the nearest-neighbour method (NN); and the prediction analysis for microarray data developed at Stanford University (PAM). NNET has allowed the distinction between different types of round blue cell tumour (3). SVM was initially used for the detection of functionally related groups of genes in *Saccharomyces cerevisiae* (13), and then for the classification of different cancer types (14). RPART analysis succeeded in detecting three genes highly related to colon cancer (11) and LDA was applied to the same colon gene dataset for tumour classification (12). Finally, the PAM methodology has been very recently employed for gene expression analysis in the small round blue cell lymphoma and in leukaemia (19). In addition, the performance of these algorithms has been evaluated considering four different techniques for reduction of data dimension. They are: the principal component analysis (24); the partial least squared analysis (25); the gene selection processes proposed in the PAM methodology (19) and in the GA/KNN technique (17). The performance of the classification techniques has been evaluated according to the percentage of misclassification by cross-validation; the lower the misclassification the better the methodology.

Robustness

We tested the robustness of the selected supervised technologies by fixing a train set of gene expression data equal to a simulated matrix with degree of confusion (CP, see Materials and Methods) of 0.5%, and then evaluating the variation of the misclassification rate on test sets with a progressively increasing degree of confusion. Figure 1 shows the results of these tests. In this analysis the misclassification rate increases up to 100% because we used matrices of data with CP greater than 50%. Actually, when the CP exceeds this threshold, the expression values of the tumour-specific upregulated genes became lower than those of the background genes, overturning the perfect classification state (see Supplementary Material for details). Our results show that SVM and NN are uniformly the most robust methods while NNET and RPART have the highest misclassification rate. LDA seems to perform at an intermediate level.

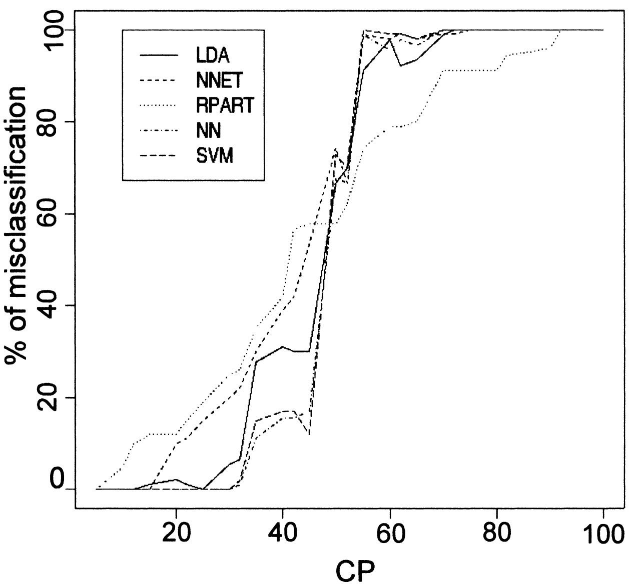

Figure 1. Comparison of the robustness of the selected supervised algorithms. Misclassification rates were compared using a train test set with CP = 0.5% and a test set with CP ranging from 0.5 to 100%. Simulation parameters (see Materials and Methods) are c and n, set equal to 3 and 90, respectively, with n_1, n_2 and n_3 all equal to 30, and $p = 30$ with p_1, p_2 and p_3 all equal to 10.

Effects of methodologies for reduction of data dimension

Next, we have examined the performances of the selected supervised technologies after the application of some methodologies for the reduction of data dimension. In particular, we have considered the principal component analysis (PC), the partial least squared analysis (PLS), the shrunken centroid technique of PAM, and the GA/KNN technique based on genetic algorithm and k-nearest neighbour (17). Figure 2 reports the most interesting results of this analysis (complete results are available in the Supplementary Material). In Figure 2A we compare the misclassification values obtained by the statistical methodologies when applied to 15 different matrices with increasing degree of confusion, with or without the aid of the PC transformation. SVM is negatively influenced by PC transformation while RPART highly improves its performance and the remaining methods do not seem to be greatly affected by PC transformation. Figure 2B shows the results of a comparison based on the same experimental approach, but in this case we used the PLS transformation as an added technology for dimension reduction. As can be seen, the PLS does not affect SVM while RPART is improving also in this case. LDA, NN and NNET seem not to be influenced by this transformation.

The effect of PLS and PC transformations on the performance of the supervised algorithms was compared using simulated matrices with (Fig. 2C) and without (Fig. 2D) 300 additional genes with randomly generated expression levels. This second series of matrices were designed to imitate real datasets of cancer gene expression where few tumour specific genes are dispersed in a large number of non-tumour-related genes. As previously demonstrated (25), we found that most algorithms (except for SVM, which prefers the PLS transformation rather than the PC) perform similarly when applied to the matrices consisting only of discriminating genes. On the other hand, we obtained different results when matrices with the 300 additional genes were analysed. LDA, RPART and SVM perform better with PLS transformation. The fitted regression lines for the three algorithms have angular coefficients of 0.67, 0.38 and 0.88, respectively. In particular, LDA becomes more efficient with high CP, while SVM improves with low CP. NNET (angular coefficient of 0.98) seems to have the same performance with PC or PLS transformations, while NN (angular coefficient of 1.1) performs better with PC transformation.

The comparison of PAM with PLS shows that LDA and RPART improve their performances with PLS (see Supplementary Material Fig. 'PAM vs PLS'). On the other hand, the comparison of PC with PAM shows that SVM has a better performance with PAM while the other techniques remain generally unchanged (see Supplementary Material Fig. 'PAM vs PC').

Finally, we studied the effect of the GA/KNN technique. The results obtained are that (i) all the classification algorithms perform similarly with GA/KNN, PC and PAM; and (ii) RPART and LDA perform much better with PLS rather than with GA/KNN. The corresponding experiments are shown in

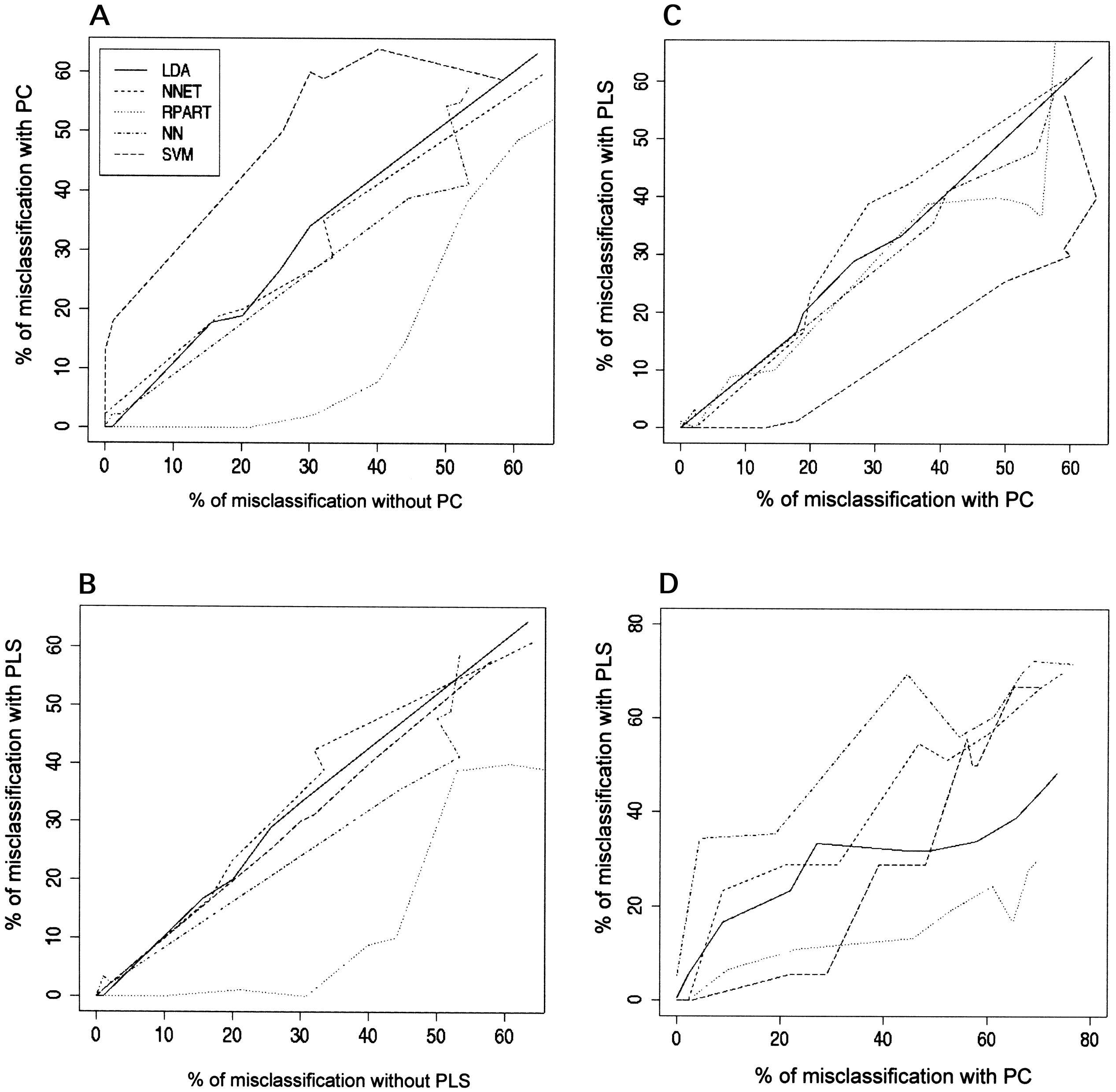

Figure 2. Scatter plots showing the misclassification rates with increasing degree of confusion using different methodologies for reduction of data dimension. (**A**) Comparison of misclassification rate with and without principal component transformation; (**B**) comparison with and without partial least square factors; (**C**) comparison of misclassification rate of PCs versus PLSs using simulated matrices that contain only discriminating genes; (**D**) comparison of misclassification rate of PCs versus PLSs using simulated matrices that are composed by the discriminating genes and 300 additional genes with random expression profiles. Simulation parameters (see Materials and Methods) in all cases are c and n equal respectively to 3 and 90, with n_1, n_2 and n_3 all equal to 30, while, $p = 30$ with p_1, p_2 and p_3 all equal to 10.

the Supplementary Material (Figs 'GA/KNN vs PC', 'GA/KNN vs PLS' and 'GA/KNN vs PAM').

Varying the number of discriminating genes, patients and class of tumours

We tested the performance of the six statistical techniques when three important experimental parameters were varied. These parameters were the number of cancer discriminating genes, the number of patients classified in a single tumour class, and the absolute number of tumour classes examined. The six graphs in the first row of Figure 3 report the misclassification trend of each statistical technique obtained by decreasing the number of discriminating genes (from 50 to 5) and by changing the degree of confusion. Different intensities of grey are representative of different percentages of misclassification (from white = 0% to black = 100%). NNET appears to be the best methodology in this analysis

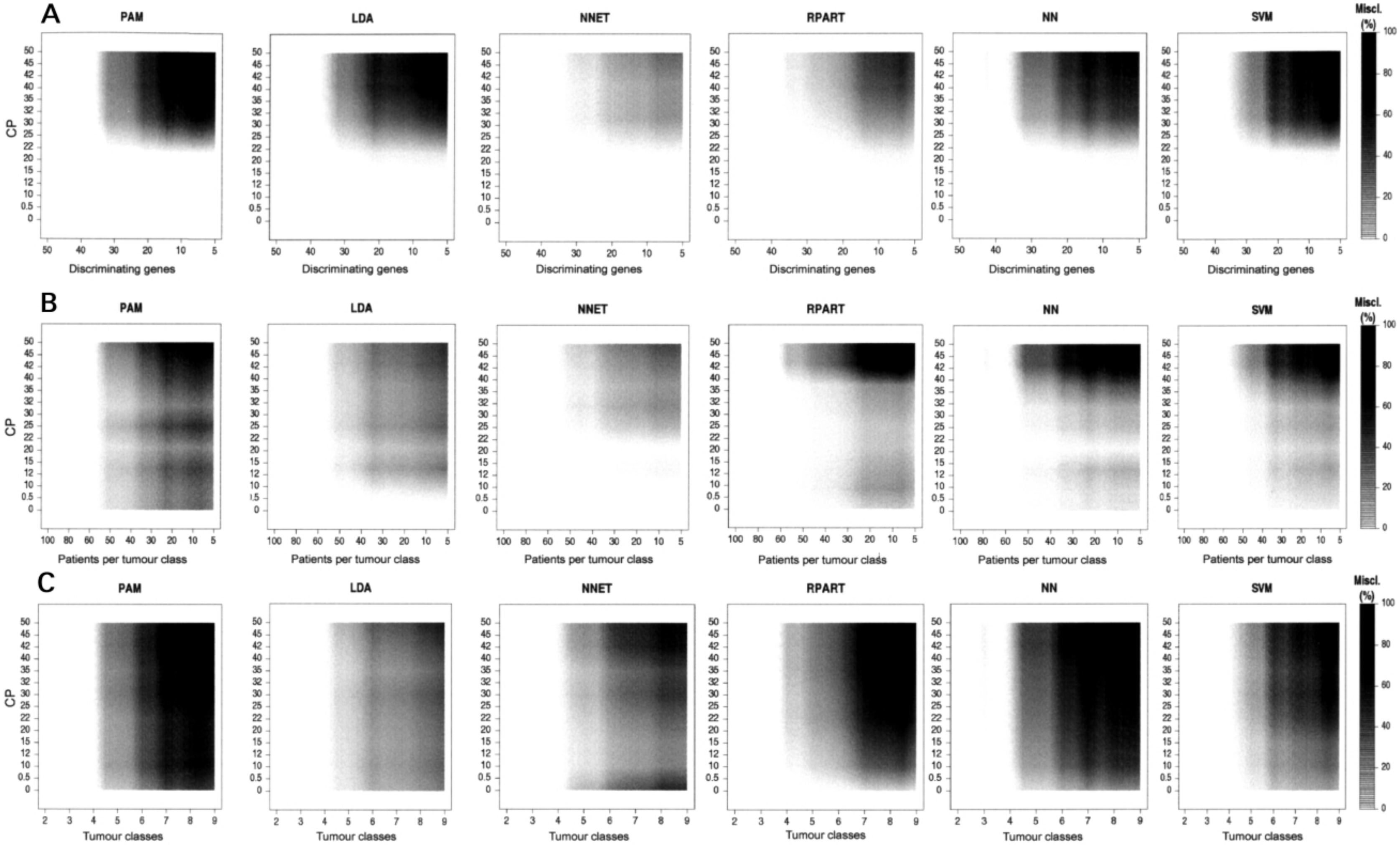

Figure 3. Misclassification rates of the six selected classification algorithms in relation to the following structural variables: (i) the decreasing number of tumour specific genes with increasing degree of confusion (diagrams of row **A**); (ii) the decreasing number of patients per tumour with increasing degree of confusion (diagrams of row **B**); and (iii) the increasing number of tumour classes with the increasing degree of confusion (diagrams of row **C**). Increasing levels of grey indicate increasing values of misclassification. This figure is presented in the Supplementary Material as a WEB tool that can be used to predict the misclassification rates of different supervised technologies in relation to the variation of the experimental parameters described above.

since it maintains a low level of misclassification with up to 30 discriminating genes per tumour class regardless the degree of confusion. When dealing with CP values lower than 15%, the NNET error rate is low even in the presence of only five discriminant genes. LDA and NN show similar classification patterns but, when the CP is higher than 15% and the number of discriminating genes is lower than 30, their misclassification rate is worse than that obtained by NNET. Under this threshold of discriminating genes, RPART, PAM and SVM methodologies have instead poorer performances and this is not dependent upon variation of CP value.

The second and the third rows of graphs in Figure 3 show the misclassification trends of the selected statistical techniques with variable number of patients per tumour class (from 100 to 5) and of tumour classes (from 2 to 9), respectively. As above, for both variables, the effect of increasing degree of confusion was simultaneously examined and different levels of grey represent different percentages of misclassification. In these tests LDA, NNET, SVM and PAM methodologies succeeded in classifying correctly most patients, regardless of the degree of confusion, with tumour class number up to 4. When this number and the CP value increase, all the methodologies become worse. Nevertheless, LDA maintains generally a better error rate. RPART and NN show instead a bad misclassification value even when they are applied to expression matrices with three classes of tumour.

A perfect classification, regardless of the CP values, is possible with a number of patients per tumour class greater or equal to 50, while with a smaller number of patients the misclassification increases, especially for high CP values. In general, NNET and LDA present the lowest misclassification rate and in particular NNET appears to be the best algorithm with CP < 20%. PAM has, in this case, the worst performance.

Analysing experimental gene expression datasets

Based on the results of the simulation approaches described above, we have tested two published datasets of gene expression profiling obtained on two different cancer types with cDNA microarrays. These are the round blue cell tumour dataset (3) and the National Cancer Institute dataset (4,5). Figure 4 and additional figures in the Supplementary Material show the misclassification trends of the statistical methodologies applied to these two selected gene expression datasets. The principal component, the partial least squares transformation, the shrunken centroid and the GA/KNN gene scoring methodology, were separately applied to the analysis of both datasets. The percentage of misclassification was then estimated with increasing number of factors (PCs and PLSs) and of genes (PAM, GA/KNN).

These tests show that the analysis of the RBC dataset achieves a general good misclassification result with all the four dimension reduction methodologies, but while with six or eight PLS factors (respectively for LDA and NNET) all the patients are correctly classified (0% error rate, Fig. 4B), with PC, shrunken centroid and GA/KNN, one patient with Ewin family tumour is misclassified as rhabdomyosarcoma (1.56% error rate, Fig. 4A and C and Fig. 'GA/KNN with RBC' in the Supplementary Material). In particular, LDA and NNET reach an error rate of 1.56% with 14 and 15 PCs, respectively (70% of variance captured). The shrunken centroid method and GA/KNN algorithm require a selection of 72 and 95 genes, respectively, to obtain the minimum error rate (for the complete list of genes see Supplementary Material). Figure 4C and F represents the misclassification trend of the PAM methodology according to an increasing set of selected genes obtained with the shrunken centroid method. As expected, the error rate increases with the decreasing number of selected genes for the classification. Furthermore, it is worth saying that RPART is the worse algorithm with both PC or with PLS transformations.

The analysis of NCI60 dataset gives very different results when PC or PLS transformations are applied. In particular, with PC all the algorithms present a general high level of misclassification (~40%). The lowest level of misclassification (30%) is reached by LDA at the 22nd PCs (80% of captured variance, see Table 1 for the confusion matrix and Fig. 4D for misclassification trend). On the other hand, with six to eight PLS factors all the tumour types are correctly recognised (0% of error rate, Fig. 4E). The shrunken centroid methodology reaches its minimum value of misclassification (28%) with a selection of more than 2000 genes. The GA/KNN technique reaches a minimum misclassification rate of 35% with 155 genes (for the list of genes see Supplementary Material). Also in this analysis the RPART shows the highest misclassification rate.

As a result of the analyses of both RBC and NCI60 datasets, we have obtained a selection of genes that show the highest factor loadings for each of the first four PLS factors and could therefore directly be involved in the biology of these tumours. These genes are listed in Tables 2 and 3, respectively.

The RBC dataset is characterized by a first factor that includes genes mostly involved in: (a) protein metabolism and modification (cathepsin B, replication protein A2 and ubiquitin-conjugating enzyme E2DE); and (b) DNA binding (such as the interleukin 3-regulated nuclear factor, NS1-associated protein 1 and inhibitor of DNA binding 3 dominant negative helix–loop–helix protein). This factor seems to be involved in the differentiation of Ewin family tumour and rhabdomyosarcoma (see Fig. 'PLS1 vs. PLS2 for RBC dataset' in the Supplementary Material). The second factor contains a variety of genes involved in cell growth and trafficking (beta actin, collagen type VII alpha 1, catenin alpha 1, cyclin A2, cadherin 3 P and osteonectin) and separates Burkitt lymphoma and Ewin family tumour from neuroblastoma and rhabdomyosarcoma. The third PLS seems to be representative of cell death and cell cycle while the fourth is representative of cell communication and DNA repair.

The analysis of the NCI60 dataset shows a less clear situation, probably related to the large number of diverse tumours that have been considered in this study. The first factor includes many genes involved in cell communication, signal transduction and immune or stress responses like the substrate of epidermal growth factor receptor kinase, the sgk gene and natural killer cells protein 4 precursor. This group of genes seems crucial in the separation of leukaemia from all other tumour types (see Figs 'PLS1 vs PLS2 for NCI60 dataset' in the Supplementary Material). The second PLS factor highly correlates with genes involved in (a) calcium, selenium or other ligand binding like S100 calcium binding protein, glutathione peroxidase 2 and insulin-like growth factor binding protein 5, and (b) transport, like caveolin-2 and glucose transporter type

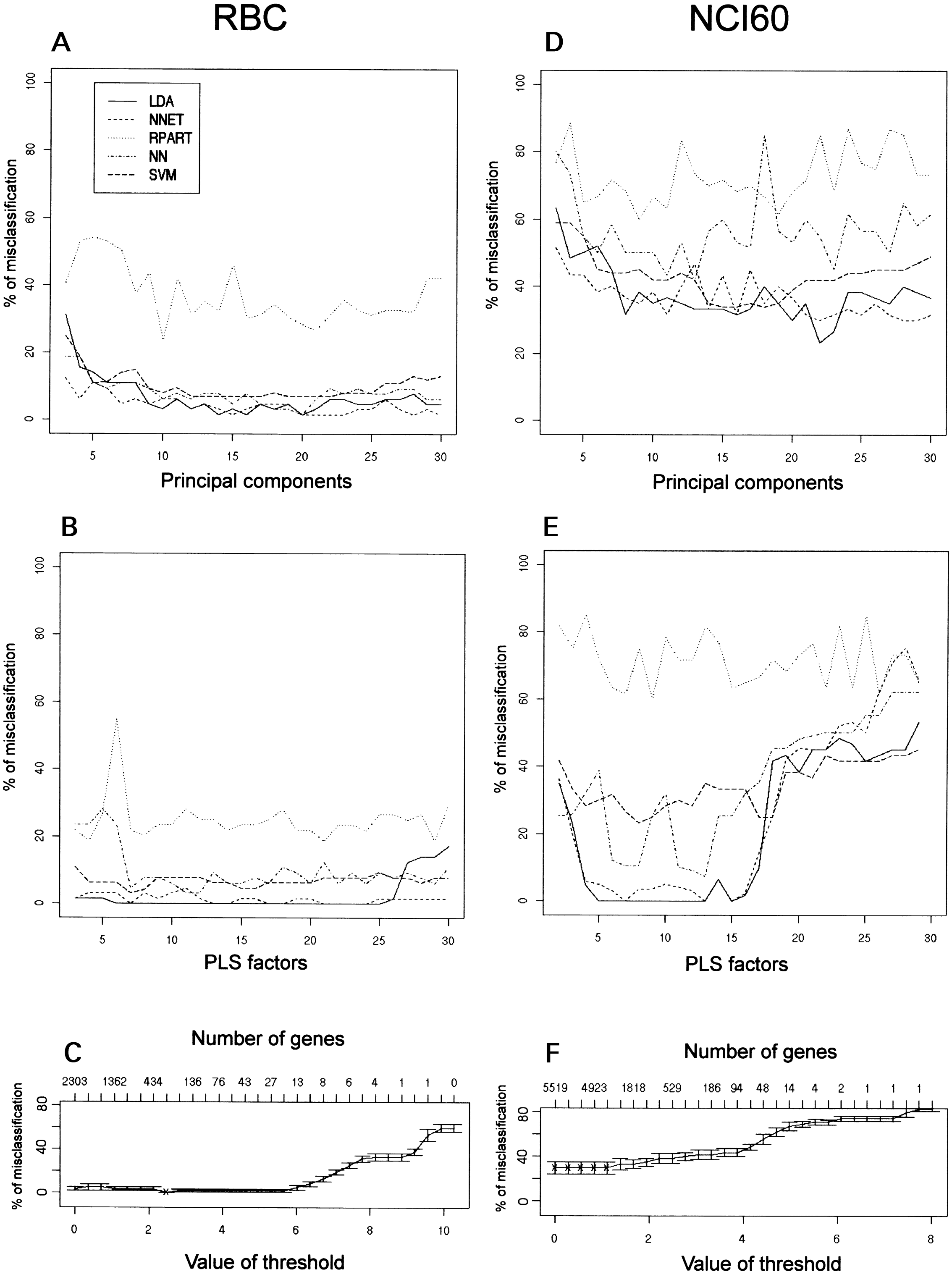

Figure 4. Comparison of the misclassification rates of the selected statistical technologies on two published gene expression datasets, with the aid of three different methodologies for reduction of data dimension. (**A**–**C**) Misclassification results (*y*-axes, misclassification percentage) of RBC dataset with increasing number of principal components, partial least square factors and genes selected with shrunken centroid method, respectively (*x*-axes). (**D**–**F**) Misclassification results of NCI60 dataset with increasing number of principal components, partial least square factors and genes selected with shrunken centroid method, respectively.

Table 1. Confusion matrix of NCI60 dataset obtained by applying the diagonalized linear discriminant analysis with 22 principal components. Italic numbers represent correctly classified patients while bold numbers represent misclassified patients

Original tumour class	Predicted tumour class								
	Melanoma	NSCLS[a]	CNS[b]	Colon	Ovarian	Renal	Breast	Prostate	Leukaemia
Melanoma	*7*	0	0	0	0	0	0	**1**	0
NSCLS	0	*4*	0	0	**2**	**1**	0	**2**	0
CNS	0	0	*5*	0	0	**1**	0	0	0
Colon	0	0	0	*5*	0	**1**	0	**1**	0
Ovarian	0	**1**	0	0	*3*	0	**1**	**1**	0
Renal	0	0	**1**	0	0	*7*	0	0	0
Breast	0	0	0	**1**	**1**	0	*5*	**1**	0
Prostate	0	**2**	0	0	0	0	0	*0*	0
Leukaemia	0	0	0	0	0	0	0	0	*6*

[a]Non-small-cell-lung carcinoma. [b]Central nervous system.

3. PLS 3 seems to be composed by genes involved in signal transduction and transport, while the fourth factor is more difficult to define since it contains genes with a variety of biological functions.

DISCUSSION

The precise classification of tumours is an issue of extreme importance for the correct diagnosis, treatment and clinical follow-up of cancer patients. In this context, DNA arrays have shown the ability to distinguish tumours that are apparently assigned to the same categories by classical clinical diagnostic approaches (26) and to predict the clinical status of cancer patients (3,6,27,28). Different statistical methodologies have been applied to the analysis of microarray data with different results (3,6,8,9,11–14,16–19,24–25). The aim of our work was the comparative evaluation of six most used statistical analyses (diagonalized linear discriminant analysis, neural networks, nearest-neighbour, support vector machine, recursive partitioning, and shrunken centroid methodology) by testing them in parallel on a series of simulated as well as on published experimental datasets and measuring their misclassification rate given by a 10-fold cross validation.

Two interdependent elements make the problem of cancer classification particularly difficult: (i) the variation of gene expression patterns among patients of the same cancer class is usually remarkable; and (ii) gene expression hallmarks of different cancer types are still not clearly defined. In this context, a simulation approach is very important since it allows the comparison through different simplified experimental situations, and this is of great help in the interpretation of the results.

Initially, the robustness of the selected discriminant techniques was compared with increasing differences between training and test sets. As expected, the non-parametric methodologies SVM and NN were demonstrated to be more robust, while RPART and NNET had the worst behaviour and LDA lay between them (Fig. 1). Secondly, the influence of dimension reduction techniques on classification results has been tested. It is well established that the performance of decision rules does not improve as the dimension of the features (genes) increases. Better results may be obtained if a feature reduction step is applied. To reduce the dimensionality of the data there are mainly two statistical approaches: the construction of new factors as linear combinations of the original variables (like the principal component analysis or partial least squared analysis) and the selection of single genes that have specific expression profile for each tumour class.

Principal component analysis reduces the high dimensionality of expression data to only few gene components, which explain as much of the observed total gene expression variation as possible, without regard to the response variable. Nguyev and Rocke compared PC dimension reduction with partial least squared analysis, mostly applied to the field of chemometrics (25). In this work the authors demonstrated that PC was competitive with PLS only if a pre-selection of the discriminating genes was performed; elsewhere PLS gave better prediction. In fact, in contrast with PC, PLS components are chosen so that the sample covariance between the response and the linear combination of genes is maximum. For our analysis we used either the principal component or partial least square analyses. As far as the single gene selection is concerned we used two recently proposed techniques for gene scoring: the shrunken centroid process of the PAM methodology (19) and the GA/KNN algorithm that is based both on a genetic algorithm and the *k*-nearest neighbour method (17).

We found that recursive partition is the only algorithm that improves its performance using either PC or PLS factors. The support vector machine, on the contrary, is badly influenced by PC transformation but not by PLS transformation. As previously demonstrated (25), we found that PLS and PC perform similarly (except for SVM) when all the genes involved in the transformation are discriminating genes. Results are different when PLS and PC are applied on the entire set of genes, including those whose expression profile is not correlated to a specific tumour category. We found that PLS can improve the classification performance of LDA, RPART and SVM techniques while not affecting NNET and worsen the results of NN (Fig. 2C and D). In particular, LDA improves its misclassification rate when dealing with a high degree of confusion, while SVM does the same with low levels of confusion.

Given the better results obtained with PLS transformation we proceeded in our simulation analysis using matrices reduced by

Table 2. Genes with the highest factor loadings in the RBC dataset are reported with their symbols and putative biological function. We have associated to each PLS factor a general biological function that summarizes the genes included in the factors. The complete list of genes is available in the Supplementary Material. Italic cells contain genes that are discussed in the text

RBC dataset	Gene symbol	Biological process
PLS 1: protein metabolism and modification—DNA binding		
Catenin (cadherin-associated protein) alpha 1 (102 kDa)	CTN1	Cell adhesion
Arylsulfatase B	ARSB	Cell organization, biogenesis
NADH dehydrogenase (ubiquinone) flavoprotein 2	NDUFV2	Electron transport
Electron-transfer-flavoprotein alpha polypeptide	ETFA	Electron transport
Recoverin	RECOV	Calcium ion binding
NS1-associated protein 1	NSAP1	RNA binding
Nuclear receptor subfamily 1 group H member 2	NRLH2	DNA binding
Nuclear factor interleukin 3 regulated	NFIL3	DNA binding
Inhibitor of DNA binding 3 dominant negative helix–loop–helix protein	ID3	DNA binding
Early growth response 1	EGR1	DNA binding
Homo sapiens HMG box containing protein 1 mRNA	HBP1	DNA binding
Cathepsin B	CTSB	Protein degradation
Insulin-like growth factor 1 receptor	IGF1R	Protein binding
Valyl-trna synthetase 2	VARS2	Protein biosynthesis
Replication protein A2 (32 kDa)	RFA2	Protein biosynthesis
Ribosomal protein S16	RPS16	Protein biosynthesis
Ubiquitin-conjugating enzyme E2D 2 (homologous to yeast UBC4/5)	UBE2DE	Protein modification
Glutamate dehydrogenase 1	GLUD1	Metabolism
Matrix metalloproteinase 2 (galatinase A)	MMP2	Extracellular matrix
PLS 2: cell growth and maintainance, calcium binding		
Epidermal growth factor receptor pathway substrate 15	EPS15	Calcium ion binding
Cadherin 3 P-cadherin (placental)	CDH3	Cell adhesion
Secreted protein acidic cysteine-rich (osteonectin)	SPARC	Cell adhesion
Metallothionein 1G	MT1G	Heavy metal binding
Hydroxyacyl-coenzyme A dehydrogenase/beta subunit	HADHB	Lipid metabolism
Adenylyl cyclase-associated protein	CAP	Cell organization, biogenesis
Arylsulfatase B	ARSB	Cell organization, biogenesis
API5-like 1	API5	Cell death
Actin beta	ACTB	Cell motility
Catenin (cadherin-associated protein) alpha 1 (102 kDa)	CTN1	Cell adhesion
Collagen type VII alpha 1	COL7A1	Cell adhesion
Erythrocyte membrane protein band 7.2 (stomatin)	EPB72	Cell shape control
Cyclin A2	CCNA2	Mitosis
Transcription factor AP-4 (activating enhancer-binding protein 4)	TFAP4	Transcription
NCK adaptor protein 1	NCK1	Signal transduction
Zinc finger protein 103	ZFP103	Nucleic acid binding
PLS 3: cell death—cell cycle		
Fas-activated serine/threonine kinase	FAST	Cell death
Msh (Drosophila) homeo box homologue 2	MSX2	Development processes
Exostoses (multiple) 1	EXT1	Development processes
Adenylyl cyclase-associated protein	CAP	Cell organization
Presenilin 1 (Alzheimer disease 3)	PSEN1	Cell death
Tumor protein p53 (Li–Fraumeni syndrome)	TP53	Cell death
Retinoblastoma-like 2 (p130)	RBL2	Cell cycle
Collagen type V alpha 2	COL5A2	Cell adhesion
Cyclin I	CCNI	Cell cycle control
Lysyl oxidase-like 2	LOXL2	Heavy metal binding
Selenium binding protein 1	SELENBP1	Selenium binding
Fatty acid binding protein 4 adipocyte	FABP4	Transport
Rhesus blood group D antigen	RHD	Transporter
Solute carrier family 16 (monocarboxylic acid transporters) member 1	SLC16A1	Transport
Membrane cofactor protein	MCP	Immune response
Coagulation factor X	F10	Haemostasis
PLS 4: cell communication and DNA repair		
Vinculin	VCL	Cell adhesion
TIA1 cytotoxic granule-associated RNA-binding protein-like 1	TIAL1	Cell death
Phosphofructokinase liver	PFKL	Energy pathway
Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta polypeptide	YWHAQ	Signal transduction
TATA box binding protein	TAF2H	Transcription regulation
Excision repair cross-complementing rodent repair deficiency complementation group	ERCC3	DNA repair
Ataxia telangiectasia mutated	ATM	DNA repair
Heterogeneous nuclear ribonucleoprotein A1	HNRPA1	Nucleic acid binding
Serine protease inhibitor Kazal type 2 (acrosin-trypsin inhibitor)	SPINK2	Immune response
Human 90-kda heat-shock protein gene	HSPCP1	Immune response
Platelet-activating factor acetylhydrolase isoform Ib gamma subunit	PAFAH1B3	Lipid catabolism

Table 3. Genes with the highest factor loadings in the NCI60 dataset are reported with their symbols and putative biological function. We have associated with each PLS factor a general biological function that summarizes the genes included in the factors. The complete list of genes is available in the Supplementary Material. Genes that are discussed in the text are in italic cells

NCI60 dataset	Gene symbol	Biological process
PLS 1: cell communication		
Epidermal growth factor receptor kinase substrate	EPS8	Signal transduction
Caldesmon	CALD1	Protein binding
Lectin, galactoside-binding, soluble, 3 (galectin 3)	LGALS3	Galactose binding lectin
Homo sapiens CAGH3 mma	CAGH3	Embryogenesis and morphogenesis
V-ets avian erythroblastosis virus E26 oncogene homolog 2	ETS2	Embryogenesis and morphogenesis
Aldehyde dehydrogenase 6	ALDH6	Metabolism/oxidoreductase
Integrin beta-5 subunit	ITGB5	Cell–matrix adhesion
Tissue inhibitor of metalloprotease 3	TIMP3	Metalloprotease inhibitor
Uncoupling protein 2	UCP2	Ion transport
Human lysophospholipase homolog	LYPLA1	Hydrolase
Homo sapiens sgk gene	SGK	Stress response
Natural killer cells protein 4 precursor	NK4	Immune response
Human BRCA2 region	BRCA2	DNA damage response
Cathepsin L	CTSL	Protein degradation
PLS 2: ligand binding and carrier transport		
S-100p protein	S100P	Calcium binding
Glutathione peroxidase 2, gastrointestinal	GPX2	Selenium binding
Insulin-like growth factor binding protein 5	IGFBP5	Protein binding
Lysyl oxidase	LOXL2	Heavy metal binding
Ets variant gene 4	ETV4	Transcription regulation
Protein-tyrosine-phosphatase	PTPRZ2	Hydrolase
Junction plakoglobin	JUP	Cell adhesion
Caveolin-2	CAV2	Protein transport
Natural killer cells protein 4 precursor	NK4	Immune response
Fibronectin 1	FN1	Immune response
Glucose transporter type 3, brain	SLC2A3	Glucose transporter
Caveolin, caveolae protein	CAV	Protein transport
PLS 3: signal transduction—transport		
Coagulation factor III	F3	Signal transducer
Basic fibroblast growth factor	FGFR1	Signal transducer
Filamin 1	FLNA	Signal transduction
Cardiac gap junction protein	GJA1	Transport
Membrane glycoprotein precursor	THY1	Protein transport
Human intestinal peptide-associated transporter hpt-1	CDH17	Protein transport
Atp synthase lipid-binding protein p1 precursor	ATP5G1	Hydrogen transport
Annexin I	ANX1	Lipid binding
Lectin, galactoside-binding, soluble, 3 (galectin 3)	LGALS3	Galactose binding lectin
Fibroblast growth factor 2	FGF2	Cell proliferation
Cysteine-rich, angiogenic inducer, 61	CYR61	Cell proliferation
Human D53 (hd53)	TPD52L1	Oncogenesis
Tissue factor pathway inhibitor 2 precursor	TFPI2	Hemostasis
PLS 4: cell communication		
Tissue inhibitor of metalloproteinase 3	TIMP3	Metalloprotease inhibitor
Glucose transporter type 3	GLUT3	Transport
Mox-2	MOX2	Transcription regulation
Urokinase-type plasminogen activator	UROK	Cell motility
Dihydrodiol dehydrogenase	DDH	Amino acid metabolism
Integrin, alpha 6	ITGA6	Cell adhesion
Laminin, alpha 3	LAMA3	Cell adhesion
Midkine (neurite growth-promoting factor 2)	MDK	Cell proliferation
Annexin I	ANX1	Lipid binding
Annexin III	ANX3	Lipid binding
Clusterin	CLU	Immune response
Glycogen phosphorylase l	PYGL	Amino acid metabolism
Glutathione s-transferase m3 (brain)	GSTM3	Glutathione transferase
Major histocompatibility complex, class II, DR beta 5	HLA-DRB5	Immune response
Human FK-506 binding protein homologue	FKBP38	Signal transduction
Tyrosine-protein kinase receptor eck precursor	TYR03	Receptor

this technique (simulation results obtained instead with principal component analysis are available in the Supplementary Material). Then, focusing on PLS transformation we compared the classification techniques along several experimental conditions to establish which methodology is most suitable in each situation. Our simulation results show that misclassification patterns are similar for all the supervised techniques but at different levels (Fig. 3). In particular, we found that most of the methodologies perform well until they are applied to a threshold number of four tumour classes, while with a higher number all the algorithms begin to misallocate individuals (Fig. 3C). LDA shows the lowest error rate even with an increasing CP degree and an increasing number of classes. Fifty homogeneous patients per tumour class seems to be the best experimental condition for classification purposes, but in the case of a low degree of confusion even smaller numbers of patients are plausible (Fig. 3B). In this case, NNET seems to perform better than all the other methods when CP is less than 22% and the number of patients is less than 50. Thirty is the minimum number of discriminating genes required for a good classification (Fig. 3A). For small CP values all the methodologies have small error rates, but NNET seems to have the lowest misclassification with $CP < 22\%$. LDA and NN show a misclassification pattern similar to the NNET one but they reach a higher error rate in worse experimental conditions.

In our test, NNET has demonstrated highly flexibility even in difficult experimental conditions such as small number of patients and discriminating genes per tumour type. Nevertheless, NNET has a poor robustness, namely it could fail to correctly classify an unknown patient case with a rather different profile from those belonging to the train set. We have used a feed-forward neural network with only one hidden layer, a very simple form of neural network. Therefore we think that NNET with different designs could improve further the ability to recognize tumour expression profiles. However it should be mentioned that neural networks could be affected by overfitting problems when they are challenged with a large amount of expression data. On the other hand, LDA showed a classification capacity similar to that of NNET but with a higher degree of robustness. The application of LDA implies the assumption of the plausible hypothesis that the expression levels within tumour types follow normal distributions. This may give an advantage to this methodology in the prediction phase; in fact non-parametric algorithms (NNET, NN, SVM and RPART) must estimate with non-parametric techniques the underlying density distribution. Non-parametric density estimation is highly influenced by the available number of observations (29) and in cancer classification the number of patients diagnosed for a certain tumour type is often quite small. This implies that the density estimation of the non-parametric classification methodologies may represent more poorly the real allocation of gene expression than normal distribution.

Two published cDNA microarray datasets were used to compare the selected methodologies and to test our simulation results. These datasets were dimensionally reduced with principal component analysis, partial least squared, shrinkage centroid and GA/KNN algorithm and then analysed with discriminant analysis on the selected features (Fig. 4 and Supplementary Material). RBC dataset shows a very small percentage of misclassification with all the dimension reduction techniques and with the majority of the classification methodologies (Fig. 4A–C and Fig. 'GA/KNN with RBC' in the Supplementary Material). NCI60 dataset has a general misclassification rate of ~40% either with PC transformation or with shrinkage centroid (Fig. 4D and F) or with GA/KNN algorithm (see figure 'GA/KNN with NCI60' in the Supplementary Material), and a perfect classification with 6 PLS factors (Fig. 4E). LDA and NNET are the best algorithms for both the datasets, but generally LDA reaches the perfect classification with the smallest number of factors.

The results obtained with these two datasets are in agreement with our simulation results. RBC has four tumour classes (situation in which all the algorithm perform well). Two of them, Ewin family tumour and rhabdomyosarcoma, contain more than 20 patients, neuroblastoma has 12 patients and Burkitt lymphoma has eight patients. Our simulation approach has shown that NNET and LDA perform better in the case of small number of patients per tumour class and this is confirmed by the analysis of the RBC expression data. The NCI60 dataset on the other hand has nine tumour classes and we have demonstrated that with this high number of classes and medium level of confusion the best technique is LDA. NCI60 has also a small number of patients per class (a minimum number of two for prostate cancer and a maximum of nine patients for non-small-cell-lung carcinoma) and NNET is the suitable algorithm in this situation.

PC and PLS factor loadings are sort of 'weights' of the original variables that contribute to the global score represented by the factor. In particular, PLS factors are constructed maximizing the covariance between the response variable (type of tumours) and the linear combination of gene expression values. Each factor should be representative of a particular feature of the data that mostly separates the different classes of tumours. Genes belonging to factors with high loadings may therefore be considered as representative of a specific tumour class. The application of the supervised and dimensional reduction technologies to both the cancer datasets, according to our experimental approach, resulted in the selection of a series of genes with the highest loadings per factor (Tables 2 and 3). In particular, proteases such as matrix metalloproteinases (MMP-2 selected in RBC and MMP-3 selected in NCI60) are capable of degrading extracellular matrix and basement membranes and have been implicated in human brain tumours. Within this group, attention has been focused on the gelatinases (MMP-2 and MMP-9), which are thought to play an important role in tumour progression (30). In the NCI60 dataset our analysis has selected fibronectin 1 that has also a matrix metalloproteinase-9 and metalloproteinase-2 secretion stimulating activity (31) and seems to play a role in tumour progression, by facilitating tissue invasion. Over expression of osteonectin (selected in the RBC dataset) in human hepatocellular carcinoma was suggested to play a role in tumour progression (32) and this gene is also expressed in human breast cancer (33). Current evidence strongly supports a role for the breast tumour suppressor genes BRCA1 and BRCA2 (found in NCI60), in both normal development and carcinogenesis. These genes have been suggested to be important for the maintenance of genome integrity and to have a role in DNA repair by homology-directed double-strand break repair (34) and mutations of BRCA2 cause familial early onset of breast and ovarian cancer. Recoverin, a retina-specific calcium-binding protein, was often found deregulated in association with small-cell lung cancer (35). The cysteine proteinases cathepsin B and L, selected by our analysis in the RBC and NCI60 datasets, respectively, are implicated in tumour invasion *in vivo* and *in vitro* as important mediators of metastasis (36). In addition cathepsin B was found overexpressed in skeletal muscle of patients with early lung cancer, and a role was suggested for this gene in inducing muscle wasting in the early stages of lung cancer (37). Reduced expression of laminin alpha 3 (selected in NCI60) and alpha 5 chains was shown in non-small cell lung cancers (38). Proteins of the cadherin family regulate cellular adhesion and motility and are believed to act as tumour suppressors. Cadherin-3, selected in the RBC dataset, was mapped close to the E-cadherin gene where frequent mutation and allelic inactivation were related to diffuse gastric cancer (39). Insulin-like growth factor binding protein, selected both in RBC and NCI60 datasets, which regulates the growth promoting effects of the IGFs on cell culture, has been reported to be expressed in poorly differentiated rhabdomyosarcoma cell lines (40). Loss of annexin 1, selected in NCI60, correlates with early onset of tumourigenesis in oesophageal and prostate carcinoma (41).

The presence of known oncogenes in the group of genes with high factors loadings makes it interesting to study the role of novel transcripts that are included in the same lists as well as of known genes for which functional evidence of direct involvement in tumours have not yet been reported.

Gene expression profiling shows great promise as a potent molecular diagnostic tool. The capacity of prediction of this approach depends on the pathologies considered, the statistical classification techniques applied for data analysis and the number of variables that are involved in each particular experimental situation. Our simulation results, confirmed by the analysis of experimental datasets of expression profiles, have demonstrated how experimental variables may affect classification results and how different classification algorithm can change their performance in relation to these variables. A better comprehension of the influence of these factors on the precision and efficacy of the results that can be obtained with the microarray technology may help the scientists initially in the phase of experimental planning and then in the phase of analysis and interpretation of expression data.

Our work proposes the method of choosing a suitable methodology for each experimental system. To this aim, a WEB tool has been constructed (available in the Supplementary Material) to help scientists who use DNA array data for cancer classification to choose the best statistical algorithms in relation to their particular experimental variables.

MATERIALS AND METHODS

Simulated matrices

The level of expression of a single gene spot in DNA microarray is represented by the non-negative ratio between two signal intensities. These signals result from the competitive hybridization for the same gene of two different RNAs (the reference and the test RNA) that are labelled with diverse fluorochromes (usually the cyanides Cy3 and Cy5). For our simulation approach, ratios are independently generated from a gamma distribution with scale parameter α set equal to 1 and shape parameter $\beta \geq 1$. Our decision of fixing the scale parameter is based on the consideration that, since mean and variance of a gamma distribution are respectively $\alpha\beta$ and $\alpha^2\beta$, the parameter α influences the simulated values mostly with an increase of variability. The increase of variability among cancer groups in our simulated matrices was performed with a shuffling of the simulated discriminating genes across tumour categories, then, we decided, for simulation strategy purposes, fixing the α parameter. Gamma distribution was already applied for microarray data modelling and testing (42), demonstrating its capability to fit well the DNA expression data. The higher the shape parameter, the higher the simulated gene expression value will be (see Supplementary Material for details on gamma distribution). Gene expression data obtained from microarray experiments can be represented as a matrix of n rows (RNA samples) and p columns (genes) where x_{ij}, with $i = 1, \ldots, n$ and $j = 1, \ldots, p$, are ratio values obtained as described before.

In our simulation, c is the number of tumour classes, $\{n_1, n_2, \ldots, n_c\}$, (with $\Sigma n_i = n$) are the number of patients per tumour and $\{p_1, p_2, \ldots, p_c\}$ (with $\Sigma p_i = p$) defines a partition of the total number of genes (see Supplementary Material for details). In a perfect classification case, the ith group of genes, p_i, will be highly expressed in only one tumour class, say i, and weakly/moderately expressed in all the other tumour classes, say j with $j = 1, 2, \ldots, c$ for all $j \neq i$. In all the simulated matrices, if not differently reported in the text, c and n are set equal to 3 and 90 respectively, with n_1, n_2 and n_3 all equal to 30 and p_i equal to 10.

Introduction of random variability

The amount of variance among tumour types and between patients with the same tumour is the cause of misclassification. Expression matrices were generated to mimic these ‘states of confusion’. The increasing degree of confusion is generated by a random introduction of variability in the expression pattern of all patients. The variability among tumours is progressively decreased and, concurrently, the variability between patients is augmented as follows: a given portion of the highly expressed, tumour-specific genes is switched off by decreasing their expression levels, while genes belonging to low expression class are switched on by increasing their expression levels. In both cases, the targets genes are randomly chosen. This portion uniquely determines the degree of confusion that hereafter will be called CP (confusion percentage). CP = 0% is the state of perfect classification while CP = 50% is the state of completely random classification.

Three hundred additional genes were introduced to simulate real cancer expression datasets where only a small proportion of genes are considered tumour-specific. Therefore, the discriminating genes represent less than the 10% of the total number of genes. The levels of expression of the extra genes were randomly generated from a gamma distribution with shape parameters randomly sampled from 1 to 10. Furthermore, the expression levels of the highly transcribed genes (p_i with $i = 1, 2, \ldots, c$) were distributed such that only 5% of the tumour-specific genes had a 10-fold level, 10% a 5-fold level, 10% a 3-fold level and 25% a 2-fold level versus the expression level of normal control.

Robustness

The performances of the selected discriminant analyses were tested in the case of training set highly different from the test set. A simulated expression matrix with $CP = 0.5\%$ (close to a situation of perfect classification) was fixed as training set while all the other matrices with $CP > 0.5\%$, were used as test sets. No transformation was applied.

Effects of methodologies for reduction of data dimension

We have generated three sets of 15 matrices only with discriminating genes and without the introduction of the 300 additional genes with random expression values. Each of the 15 matrices corresponds to a different level of confusion ranging from 0 to 50%. The first set has original simulated expression levels, the second was equal to the first set but with matrices converted into principal components (PC), the third was equal to the first but with matrices converted into factor obtained with partial least squared analysis (PLS). Then, we compared the misclassification results at increasing degree of confusion: PLS versus no transformation, PC versus no transformation, and PLS versus PC. Subsequently, PLS and PC transformations were also confronted on two other sets of matrices containing the 300 additional genes. Furthermore, shrunken centroid and GA/KNN technique were applied on the simulated matrices and genes with the highest scoring levels were selected. Then, discriminant techniques were applied only on these sets of genes. Misclassification results were compared with those of PLS, PC.

Decreasing the number of discriminating genes

The performances of the six selected statistical methodologies were analysed according to the variation of the number of tumour-specific genes. After fixing the number of tumour classes (c) to 3 and the number of patients per tumour class (n_i with $i = 1, 2, 3$) to 30, we generated matrices with equally decreasing p_1, p_2 and p_3 from 50 to 3 for CP values ranging from 0 to 50%. A total of 90 matrices were generated.

Increasing the number of tumour classes and patients per class

The performances of the six selected statistical techniques were compared according to the variation of the number of tumour class and of the number of patients per tumour class. For each CP value, we created eight different matrices with a fixed number of 30 patients per tumour class and with increasing number of tumour class from 2 to 9. A total number of 120 matrices were produced. Furthermore, for each CP value, nine matrices were generated with numbers of homogeneous patients per tumour ($n_1, n_2, \ldots, n_c$) decreasing from 100 to 5, while the number of tumour classes was maintained equal to 3. This resulted in the production of a total of 135 additional matrices.

Published datasets

Two experimental gene expression datasets were used to compare the selected discriminant analyses: (i) the NCI60 dataset—in a project of the National Cancer Institute for the screening of anti-cancer drugs, a cDNA array with 5519 genes was used to study the expression profiles of 60 human cell lines derived from nine tumour classes (4–5); (ii) the RBC dataset—a cDNA array of 2303 genes was used to study the expression signatures of four round blue cell tumours of childhood (3). These datasets are available at http://genome-www.stanford.edu/nci60 (NCI60) and at www.nature.com (RBC) respectively.

Biological interpretation of factors with published datasets

For each expression dataset we calculated the number of PLS factors needed to reach the minimum rate of misclassification. For each factor we then identified a list of genes with the highest PLS loadings that can be considered as markers of each class of tumour. All the genes were annotated with their correspondent biological function according to the AmiGO browser for gene ontology (www.geneontology.org). Each gene participates, albeit at different degrees, in all the extracted components. This means that a single gene can be involved in more than one functional network characteristic of a different class of cancer.

Statistical techniques

Six different supervised learning techniques were chosen for our comparison (43,44): (i) the diagonalized linear discriminant analysis (LDA); (ii) the recursive partitioning (RPART); (iii) the feed-forward neural networks (NNET) with only one hidden layer; (iv) the support vector machine (SVM); (v) the nearest neighbour (NN); and (vi) the shrunken centroid methodology (PAM).

Furthermore, four dimension-reduction techniques were chosen for the analysis: principal component analysis, partial least squared analysis, the shrunken centroid gene scoring methodology implemented in an freely available software for R package, called PAM (www-stat.stanford.edu/~tibs/PAM/index.html) and the GA/KNN algorithm kindly provided by the authors. For a more detailed description of these statistical techniques see the Supplementary Material. The analyses described in the Results section were performed with dedicated functions implemented for the statistical package R for Linux system, available at www.r-project.org (43,44).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online and also at http://muscle.cribi.unipd.it/microarrays/simulation/.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Mike Denham of the Department of Applied Statistic, University of Reading, UK for providing R functions for partial least squares analysis and Dr Leping Li and his colleagues for kindly providing the source code of the algorithm GA/KNN. This work was supported by the Fondazione Telethon ONLUS, Italy (grant no. B57 to G.V. and G.L.), by the Ministero dell'Università e della Ricerca Scientifica, Italy (grant COFIN 2000 no. MM05192239 and COFIN 2001 to G.L.) and by the Ministero Della Sanità, Italy (Progetto Finalizzato ICS 060.1/KF 99-77 to G.L.).

REFERENCES

1. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

2. Wooster, R. (2000) Cancer classification with DNA microarray. *Trends Genet.*, **16**, 327–329.
3. 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. *Nat. Med.*, **7**, 673–679.
4. Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
5. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
6. Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S. *et al.* (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, **8**, 68–74.
7. Hanahan, D. and Weinberg, R. (2000) The hallmark of cancer. *Cell*, **100**, 57–71.
8. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression pattern. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
9. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps, methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
10. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000) Tissue classification with gene expression profiles. *J. Comput. Biol.*, **7**, 559–583.
11. Zhang, H., Yu, C.Y., Singer, B. and Xiong, M. (2001) Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 6730–6735.
12. Xiong, M., Jin, L., Li, W. and Boerwinkle, E. (2000) Computational methods for gene expression-based tumor classification. *Biotechniques*, **29**, 1264–1268.
13. Brown, M.P., 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. Natl Acad. Sci. USA*, **97**, 262–267.
14. Furey, T.S, Cristianini, N., Duffy, N., Bednarski, D.W., Shummer, M. and Haussler, D. (2000) Support vector machine and validation of cancer tissue using microarray expression data. *Bioinformatics*, **16**, 906–914.
15. McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York.
16. 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**, 239–247.
17. Li, L., Weinberg, C.R., Darden, T.A. and Pedersen, L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.
18. Landgrebe, J., Wurst, W. and Welzl, G. (2002) Permutation-validated principal components analysis of microarray data. *Genome Biol.*, **3**, research0019.1–0019.11.
19. Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.
20. Dudoit, S., Fridlyand, J. and Speed, T. (2000) Comparison of discrimination methods for the classification of tumours by using gene expression data. *J. Am. Stat. Assoc.*, **97**(457), 77–87.
21. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallom M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
22. Crescenzi, M. and Giuliani, A. (2001) The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *FEBS Lett.*, **507**, 114–118.
23. Benigni, R. and Giuliani, A. (1994) Quantitative modeling and biology, the multivariate approach. *Am. J. Physiol.*, **266**, 1697–1704.
24. Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000) Principal component analysis to summarize microarray experiment, application to sporulation time series. *Pac. Symp. Biocomput.*, **2000**, 455–466.
25. Nguyen, D.V. and Rocke, D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.
26. Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Behesti, J., Bueno, R., Gillette, M. *et al.* (2001) Classification of human lung carcinoma by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.
27. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R. and Nevins, J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.
28. Yeang, C.H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R.M., Angelo, M., Reich, M. and Lander, E. (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17**, 316–322.
29. Silverman, B.W. (1986) *Density Estimation for Statistic and Data Analysis*. Chapman and Hall, Bristol.
30. Rooprai, H.K. and McCormick, D. (1997) Proteases and their inhibitors in human brain tumours, a review. *Anticancer Res.*, **17**, 4151–4162.
31. Thant, A.A., Nawa, A., Kikkawa, F., Ichigotani, Y., Zhang, Y., Sein, T.T., Amin, A.R. and Hamaguchi, M. (2000) Fibronectin activates matrix metalloproteinase-9 secretion via the MEK1-MAPK and the PI3K-Akt pathways in ovarian cancer cells. *Clin. Exp. Metastasis*, **18**, 423–428.
32. Le Bail, B., Faouzi, S., Boussarie, L., Guirouilh, J., Blanc, J.F., Carles, J., Bioulac-Sage, P., Balabaud, C. and Rosenbaum, J. (1999) Osteonectin/SPARC is overexpressed in human hepatocellular carcinoma. *J. Pathol.*, **189**, 46–52.
33. Bellahcene, A. and Castronovo, V. (1997) Expression of bone matrix proteins in human breast cancer, potential roles in microcalcification formation and in the genesis of bone metastases. *Bull. Cancer*, **84**, 17–24.
34. Tutt, A., Bertwistle, D., Valentine, J., Gabriel, A., Swift, S., Ross, G., Griffin, C., Thacker, J. and Ashworth, A. (2001) Mutation in Brca2 stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. *EMBO J.*, **20**, 4704–4716.
35. Bazhin, A.V., Shifrina, O.N., Savchenko, M.S., Tikhomirova, N.K., Goncharskaia, M.A., Gorbunova, V.A., Senin, I.I., Chuchalin, A.G. and Philippov, P.P. (2001) Low titre autoantibodies against recoverin in sera of patients with small cell lung cancer but without a loss of vision. *Lung Cancer*, **34**, 99–104.
36. Berndt, A., Johannesson, J., Haas, M., Arkona, C., Katenkamp, D., Wiederanders, B., Hyckel, P. and Kosmehl, H. (2000) Simultaneous enzyme overlay membrane (EOM)-based *in situ* zymography and immunofluorescence technique reveals cathepsin B-like activity in a subset of tumour vessels. *Histochem. Cell. Biol.*, **114**, 63–68.
37. Jagoe, R.T., Redfern, C.P., Roberts, R.G., Gibson, G.J., Goodship, T.H. (2002) Skeletal muscle mRNA levels for cathepsin B, but not components of the ubiquitin-proteasome pathway, are increased in patients with lung cancer referred for thoracotomy. *Clin. Sci.*, **102**, 353–361.
38. Akashi, T., Ito, E., Eishi, Y., Koike, M., Nakamura, K. and Burgeson. R.E. (2001) Reduced expression of laminin alpha 3 and alpha 5 chains in non-small cell lung cancers. *Jpn. J. Cancer Res.*, **92**, 293–301.
39. Braungart, E., Schumacher, C., Hartmann, E., Nekarda, H., Becker, K.F., Hofler, H., Atkinson, M.J. (1999) Functional loss of E-cadherin and cadherin-11 alleles on chromosome 16q22 in colonic cancer. *J. Pathol.*, **187**, 530–534.
40. Ayalon, D., Glaser, T. and Werner, H. (2001) Transcriptional regulation of IGF-I receptor gene expression by the PAX3-FKHR oncoprotein. *Growth Horm. IGF Res.*, **11**, 289–297.
41. Paweletz, C.P., Ornstein, D.K., Roth, M.J., Bichsel, V.E., Gillespie, J.W., Calvert, V.S., Vocke, C.D., Hewitt, S.M., Duray, P.H., Herring, J. *et al.* (2000) Loss of annexin 1 correlates with early onset of tumorigenesis in esophageal and prostate carcinoma. *Cancer. Res.*, **60**, 6293–6297.
42. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blatter, F.R. and Tsui, K.W. (2001) On differential variability of expression ratios, improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.*, **8**, 37–52.
43. Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
44. Venables, W.N. and Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS*, 3rd edn. Springer, Berlin.