BIOINFORMATICS

Vol. 19 no. 5 2003, pages 571–578
DOI: 10.1093/bioinformatics/btg051

PCA disjoint models for multiclass cancer analysis using gene expression data

S. Bicciato, A. Luchini and C. Di Bello*

Department of Chemical Process Engineering, University of Padova, via Marzolo, 9, 35131, Padova, Italy

Received on May 27, 2002; revised on October 7, 2002; accepted on October 30, 2002

ABSTRACT

Motivation: Microarray expression profiling appears particularly promising for a deeper understanding of cancer biology and to identify molecular signatures supporting the histological classification schemes of neoplastic specimens. However, molecular diagnostics based on microarray data presents major challenges due to the overwhelming number of variables and the complex, multiclass nature of tumor samples. Thus, the development of marker selection methods, that allow the identification of those genes that are most likely to confer high classification accuracy of multiple tumor types, and of multiclass classification schemes is of paramount importance.

Results: A computational procedure for marker identification and for classification of multiclass gene expression data through the application of disjoint principal component models is described. The identified features represent a rational and dimensionally reduced base for understanding the basic biology of diseases, defining targets for therapeutic intervention, and developing diagnostic tools for the identification and classification of multiple pathological states. The method has been tested on different microarray data sets obtained from various human tumor samples. The results demonstrate that this procedure allows the identification of specific phenotype markers and can classify previously unseen instances in the presence of multiple classes.

Availability: Matlab source codes are available from the authors.

Contact: silvio.bicciato@unipd.it

Supplementary information: http://www.dpci.unipd.it/PersPages/SBicciato/SIMCArray.html

INTRODUCTION

DNA microarrays are radically boosting the understanding of living systems, thus creating enormous opportunities to elucidate the biological processes of cells in different physiological states. In particular, the application of high-throughput technologies to analyze DNA, RNA or proteins from tumor cells is improving cancer analysis to levels that classical methods have been unable to reach. Several studies provide clear examples of how molecular forecasting of cancer outcome represents a significant adjunct to existing prognostic methods (Golub *et al.*, 1999; Alon *et al.*, 1999; Perou *et al.*, 1999; Alizadeh *et al.*, 2000; Bittner *et al.*, 2000; Khan *et al.*, 2001; Ramaswamy *et al.*, 2001; Armstrong *et al.*, 2002). However, cancer analysis and classification on the basis of microarray data poses the challenge to develop computational procedures able to address specific issues, such as modeling multiple, heterogeneous populations and reducing the overwhelming number of variables (genes).

In particular, the presence of samples belonging to multiple categories hampers the development of procedures for the molecular classification of distinct tumor types. Indeed, most of the proposed methods tackle only binary classification problems or are difficult to extend into multiclass versions (Yeang *et al.*, 2001). The problem is further complicated by the fact that most of the monitored expression profiles may not be relevant to the description of the pathological state. As such, these variables could potentially degrade the performance of the classification scheme by masking the contribution of the relevant features. Thus, together with the development of classification tools in the context of multiple tumor types, the identification of those genes that are most likely to confer high classification accuracy is of paramount importance.

Two major approaches have been adopted to develop multiclass classification procedures; namely, the direct application of a multiclass model (i.e. neural networks in Khan *et al.*, 2001) and the combination of several binary classifiers in conjunction with different multiclass prediction schemes (Yeang *et al.*, 2001; Ramaswamy *et al.*, 2001; Pomeroy *et al.*, 2002).

The purpose of this work is to present a multivariate procedure that allows: (i) *marker identification* by extracting genes related to specific pathological states; and (ii) *robust diagnosis* by accurately predicting the class of unlabeled tumor samples from the gene expression profiles of mul-

*To whom correspondence should be addressed.

tiple tumor types. Principal component analysis (PCA) is used to implement the modeling scheme called Soft Independent Modeling of Class Analogy (*SIMCA*), originally developed by Wold (Wold, 1976). In a multiclass problem, SIMCA works by considering each class separately. For each class, a principal component analysis is performed leading to a different PCA model for each category (thus called *disjoint class models*). Since the models are disjoint, the system describing one class does not depend on that of another category. When classifications of unknown samples are attempted, a comparison is made between the unclassified sample and each class model. Class assignment is achieved by finding the model that best fits the unknown specimen within a specified statistical significance. Even if reliable classification of previously unseen instances is the ultimate goal of the original approach, SIMCA has been adapted to solve the fundamental issue of feature selection in the context of multiclass tumor analysis using gene expression data. Thus, examining the structure of the variance explained by each model, it is possible to distinguish among the most important variables characterizing each single class and identify specific genes most highly correlated with the different tumor types.

This multiclass modeling approach has been applied to two gene expression databases involving various human tumor classes: (1) the data set from Golub *et al.* (1999) on acute leukemia classification; and (2) the study presented by Khan (Khan *et al.*, 2001) on small round blue-cell tumors.

In this paper, the Methods section describes the implementation of the disjoint principal component scheme, the SIMCA classification rule, and the feature selection procedure based on the *classifier feedback* method. The Results section presents the application of the proposed approach to the analysis of gene expression data. Additional tables and figures are available in the Supplementary information section (denoted as $\_SI$ throughout the test). The Results section also includes permutation analysis to examine the reproducibility and stability of the identified markers. The final Conclusion section discusses the proposed approach in comparison with other methods for the analysis of multiclass gene expression profiles and highlights future developments of the current project.

METHODS

SIMCA modeling technique exploits the properties of principal components analysis to extract patterns from a set of objects. These characteristics are then used to analyze the different classes of the data set and assign previously unseen objects to the class they resemble the most. The method, developed by Wold (Wold, 1976), assumes that the objects in a single separate class are in some way similar. On the basis of this similarity, a principal component model is formulated on the objects defining each single class. The total model for a multiclass system consequently consists of a collection of disjoint PCA models, one for each class. Unclassified samples are then fitted to all calibrated class models and classified as belonging to the model they statistically best fit. As such, SIMCA also accounts for the possibility that unclassified objects might define a new class, not fitting any of the calibrated models, or might resemble the characteristics of more than one class. Even if it has been mostly applied to solve classification problems, SIMCA can also be used for feature selection, meaning the identification of those discriminating variables that better characterize a category.

Modeling scheme

A training matrix $\mathbf{X}$ consists of n objects from Q different known classes described by m variables. The observations $x_{ki}^{(q)}$ of any submatrix $\mathbf{X}^{(q)}$ $(n_q \times m)$, containing $n_q < n$ training objects belonging to class q, are modeled separately by PCA. PCA is a widely used data analysis technique that allows reducing the dimensionality of the system while preserving information on the variable interactions (Joliffe, 1986). PCA transforms the original variables into a set of linear combinations, the principal components (PC), which capture the data variability, are linearly independent and weighted in decreasing order of variance coverage. This allows a straightforward reduction of the data dimensionality by discarding the feature elements with low variability. Thus, all original *m-dimensional* data patterns can be optimally transformed to data patterns in a feature space with lower dimensionality. Singular value decomposition (SVD), the algorithm used in this work, or other decomposition methods can be applied to perform PCA after column autoscaling. Theoretical aspects of PC calculation will be omitted since several texts address the topic in detail (e.g. Joliffe, 1986). After calibrating a principal component model for each class separately, data are described by a number of disjoint models:

$$z_{ki}^{(q)} = \sum_{a=1}^{A_q} t_{ka}^{(q)} l_{ai}^{(q)} + e_{ki}^{(q)} \qquad (1)$$

where $z_{ki}^{(q)}$ are the autoscaled training data of class q, A_q the number of significant principal components in class q, $l_{ai}^{(q)}$ the loading on component a of variable i in class q, $t_{ka}^{(q)}$ the score of object k on component a in class q, and $e_{ki}^{(q)}$ the residual of object k of the class q training set at variable i. This is equivalent to any PCA analysis, with the additional symbol q indicating that only the training data of class q are considered to construct the model.

The number A_q of significant principal components can be obtained using a cross-validation technique or setting a threshold on the minimum variance explained by A_q factors.

Once the PC models of the Q classes have been derived, a *class region* is constructed around the A_q PCs calculating the residual standard deviation for each class, $s_0^{(q)2}$, from the residuals $e_{ki}^{(q)}$ of the class q training samples:

$$s_0^{(q)2} = \sum_{k=1}^{n_q} \sum_{i=1}^{m} (e_{ki}^{(q)})^2 / [(n_q - A_q - 1)(b - A_q)] \quad (2)$$

In Equation (2) the denominator represents the degrees of freedom computed according to De Maesschalck *et al.* (1999) and b is the minimum between $n_q - 1$ and m.

The observations of any unclassified object p are then fit to all the Q models (1) with the same values of the autoscaling and $l_{ai}^{(q)}$ parameters. The variance of the deviations $s_p^{(q)2}$ indicates how well object p fits class q:

$$s_p^{(q)2} = \sum_{i=1}^{m} (e_{pi}^{(q)})^2 / (b - A_q) \quad (3)$$

Classification rule

The final classification of object p is obtained comparing its residual variances to the residual variance within each class q through an F-test:

$$F = \frac{s_p^{(q)2}}{s_0^{(q)2}} \quad (4)$$

If the F-value is smaller than the critical F-value (F_{limit}) at a given level of significance (i.e. 0.05) for $(b - A_q)$ and $(n_q - A_q - 1)(b - A_q)$ degrees of freedom, object p can be assigned to class q. It should be emphasized that, given the rule of Equation (4), there are 3 possible classification outcomes: (1) sample is exclusively assigned to one class; (2) sample does not belong to any class; (3) sample belongs to 2 or more classes. The second category is the result of a poor fit of the sample to the existing models. In this case, the sample may represent an *outlier* or an individual of a previously unconsidered new population and is labeled as *neither*. The third fate arises if the original classes do not contain specific information to be statistically different and the sample will be labeled as *multiple*.

Further details of the SIMCA procedure can be found in Wold (1976); Massart *et al.* (1988), and De Maesschalck *et al.* (1999).

Feature selection

Feature selection is usually defined as the process of finding a subset of characteristics, from the original set of variables, optimal according to a defined goal or selection criterion. One of the selection paradigms indicates to select a feature subset that guarantees maximal *within-class modeling power* and *between-class separability*. This process helps the design of optimal predictor-classifiers exploiting ability of features to reproduce and distinguish patterns from different classes. Among the existing feature selection methods, the procedure used in this work is a model-dependent analysis, also known as *classifier feedback* approach (John *et al.*, 1994). According to this method, the quality of a selected feature subset is evaluated using as a criterion the performance of the classification algorithm for the reduced data set.

The feature selection procedure comprises three major steps: (i) identification of those variables that best describe any given class (i.e. the creation of class-specific lists of genes based on the modeling power of the original variables); (ii) scoring and ranking of the variables in each class-related list according to their ability to discriminate the class they model from all the other categories; and (iii) computation of the minimum number of variables needed to maximize multiclass classification.

The Q models defined in Equation (1) are used to sort and rank the different descriptors of the system in terms of their ability to describe a specific category while discriminating among the different classes. A *class-q-variable* is defined so that it presents large values of the residuals when *class-q-samples* are fitted to all categories but the true q model and, at the same time, the error of the q model is minimized only by *class-q-samples*. Specifically, for each of the original m variables, Q relevance indexes $r_i^{(q)}$ are calculated as the weighted power of a variable to model the training objects of class q compared to the training samples of all the other $Q - 1$ classes:

$$r_i^{(q)} = \frac{\sum_{r=1;r\neq q}^{Q-1} \left(\sum_{k=1}^{n_r} (e_{ki}^{(q)})^2 \right)_r \Big/ (Q-1)}{\frac{\sum_{k=1}^{n_q} (e_{ki}^{(q)})^2}{n_q \sigma_i^{(q)2}}} \quad (5)$$

In Equation (5), the numerator represents the average error generated by variable i when the samples of all $Q - 1$ classes different from q are projected into model q. The median can replace the mean to smooth the effect of possible *outlying* samples or if the number of samples in the training sets is relatively small.

Given Equation (5), each variable i is characterized by Q values of $r_i^{(q)}$ and is assigned to the class q for which its $r_i^{(q)}$ value is higher (*modeling power*). Being this first step a *winner-take-all* exclusive assignment, the original m variables are partitioned into Q disjoint subsets $M^{(q)}$, composed of $m^{(q)}$ unique putative features. However, when a large number of noisy variables are

considered, some of the elements in the $M^{(q)}$ subsets may present similar $r_i^{(q)}$ values and, while assigned to different classes, variables with comparable $r_i^{(q)}$ may not be able to uniquely describe a single category. Therefore, a further step is needed to refine $M^{(q)}$ lists and sort out the non-discriminating descriptors.

In the second step, variables inside any $M^{(q)}$ partition are ranked based on their *discriminatory power*, or the ability to maximize not only the average error generated when projecting all classes into model q, but also the residual standard deviation of every single class. Specifically, each variable is scored according to the error induced by any single training set different from q (i.e. $\sum_{k=1}^{n_r}(e_{ki}^{(q)})^2$ with $r \neq q$) and its final rank is computed as the average of the scores obtained by each $m^{(q)}$ variable over the $Q-1$ training sets different from q.

Once defined Q sorted lists from the original m variables, a systematic procedure for determining the minimum number of features needed to describe a phenotype and maximize multiclass classification is applied. The feature selection adopts SIMCA classification performance on a previously unseen set of objects (*test set*) as the search criterion. Several SIMCA models are calibrated on the objects of the Q training sets using different numbers of variables starting from the top-ranking in each sorted list (e.g. the first 5, 10, 20, etc. top-ranking variables in each class-list). For each block of class variables, the percentage of correct classification over the test set objects is calculated and the group of selected features is defined as the subset of class variables that maximizes the classifier performance (*class markers*).

RESULTS

The procedure based on PCA disjoint class models has been applied to the analysis of two gene expression databases involving various human tumor classes, namely the data set from Golub *et al.* (1999) on acute leukemia classification and the study presented by Khan *et al.* (2001) on small round blue-cell tumors. In both cases, the goal of the computational approach has been the selection of peculiar class markers and the assignment of previously unseen samples to multiple tumor classes.

Acute leukemia

The acute leukemia study provides measurements for 7129 probes in 72 samples collected from acute leukemia patients with 47 cases diagnosed as acute lymphoblastic leukemia (ALL) and the other 25, as acute myeloid leukemia (AML). Following the experimental setup described in Golub *et al.* (1999), data has been split into a training set of 38 samples (19 B- ALL, 8 T-ALL, and 11 AML) and a blind test set of 34 samples (19 B-ALL, 1 T-ALL, and 14 AML). Before the application of the SIMCA scheme, gene expression values have been subjected to a variation filter that excluded genes showing minimal variation across the samples being analyzed and reduced the number of variables to 3930 (Armstrong *et al.*, 2002).

With the aim to first quantify the relative relevance of each transcript in describing different subtypes of leukemia, three PCA models are built using ALL-B, ALL-T, and AML training samples. A total of 4, 4, and 2 principal components accounting for the 71.9, 73.2, and 88.5% of the overall variance, respectively, describe the global model (Table 1_SI). The number of principal components has been determined using a leave-one-out cross-validation procedure (details described in Joliffe, 1986). The smaller number of principal components needed to describe the AML class indicates that AML objects are characterized by a peculiar multivariate structure.

The analysis of residuals in the three models allowed partitioning the original 3930 transcripts in 615 related to the class ALL-B, 2657 to the class ALL-T, and 658 to AML. After ranking the variables in the 3 sorted lists, several SIMCA models have been calibrated on the training sets to determine the minimum number of features needed to describe each phenotype. Specifically, 76 different SIMCA classifiers are trained using variable subsets of increasing size created from the sorted list of each class. As an example, in Tables 2_SI and 3_SI the subset of 6 total variables is composed selecting the first 2 of the 615 ALL-B, the first 2 of the 2657 ALL-T, and the first 2 of the 658 AML class-specific genes. For comparison, a SIMCA classifier has been calibrated also using all 3930 original variables. In this phase, the number of principal components has been automatically determined setting a threshold of 70% on the minimum variance explained. For any variable subset, the classification of the 34 test set objects has been calculated using SIMCA classification rule. Figure 1a and Table 3_SI show the classification performance in terms of correctly classified, misclassified, multiple and neither test samples for the different variable subsets. The best classification performance, defined as the higher percentage of correct assignments over the test set objects, is obtained using the top $30 \div 32$ variables of each class subset. In particular with 30 variables/class, 28 samples have been correctly classified, 2 samples have been recognized by more than one model and 4 samples did not statistically resemble the characteristics of any training set (Tables 4_SI and 5_SI). The best test set performance (i.e. 82.3% of correctly identified samples) could seem lower than the classification results obtained by Golub (Golub *et al.*, 1999), Chow (Chow *et al.*, 2001), or Nguyen (Nguyen and Rocke, 2002) using other modeling schemes. However, it has to be noted that

no sample has been misclassified and that the relatively poorer classification performance is only determined by the presence of samples labeled as *multiple* and *neither*. Specifically, the specimen labeled as *multiple* are #66 and #67, two samples that are not properly assigned also using other classification methods (Golub *et al.*, 1999; Nguyen and Rocke, 2002). Instead, the *neither* category accounts for those samples that cannot receive a statistically confident classification and is comparable to the *Prediction Strength* defined by Golub *et al.* (1999) or the *Diagnosis index* introduced by Khan *et al.* (2001). Indeed, as shown in Table 6_SI, samples labeled as *neither* can receive a correct classification considering only their *distance* from the three models (described by the variance of the deviations $s_p^{(q)2}$), but this call cannot be considered statistically confident.

Table 7_SI lists the top-30 features (*class markers*) that maximize the classifier performance while Figures 1_SI, 2_SI, and 3_SI show the expression levels in the three classes for some of the identified markers. Raw data, lists of top-discriminators from previous studies (Golub *et al.*, 1999; Keller *et al.*, 2000; Chow *et al.*, 2001), and experimental evidences support the specificity of the selected features (Table 7_SI).

Small, round blue-cell tumors

The small, round blue-cell tumors (SRBCT) data set consists of 2308 gene-expression profiles from cDNA experiments describing four childhood malignancies: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS). The original training set consists of 63 samples and includes both tumor biopsy material and cell lines. The independent test set consists of 25 samples representing tumors, cell lines, and non-SRBCT specimen (2 normal muscle tissues, Test 9 and Test 13, and 3 cell lines, Test 3, Test 5, and Test 11).

Four PCA disjoint models have been built using the training samples of each class (Table 8_SI). In the global model, 8 components model EWS samples (explaining 73.2% of the overall variance), 8 components are needed for RMS (72.5% of the total variance), and 6 and 4 principal components describe NB and BL classes, respectively (75.6 and 73.7% of the overall variance). The analysis of the model residuals allowed partitioning the original 2308 variables in four sorted subsets of 600, 496, 512, and 700 genes related to EWS, RMS, NB, and BL respectively. To identify the transcripts characterizing each SRBCT class, SIMCA modeling scheme has been applied to the independent test set using different subsets of variables starting from the top-ranking in any sorted list. Specifically, 90 different SIMCA classifiers have been calibrated using different gene subsets and all 2308 original variables (Table 9_SI). Figure 1b and

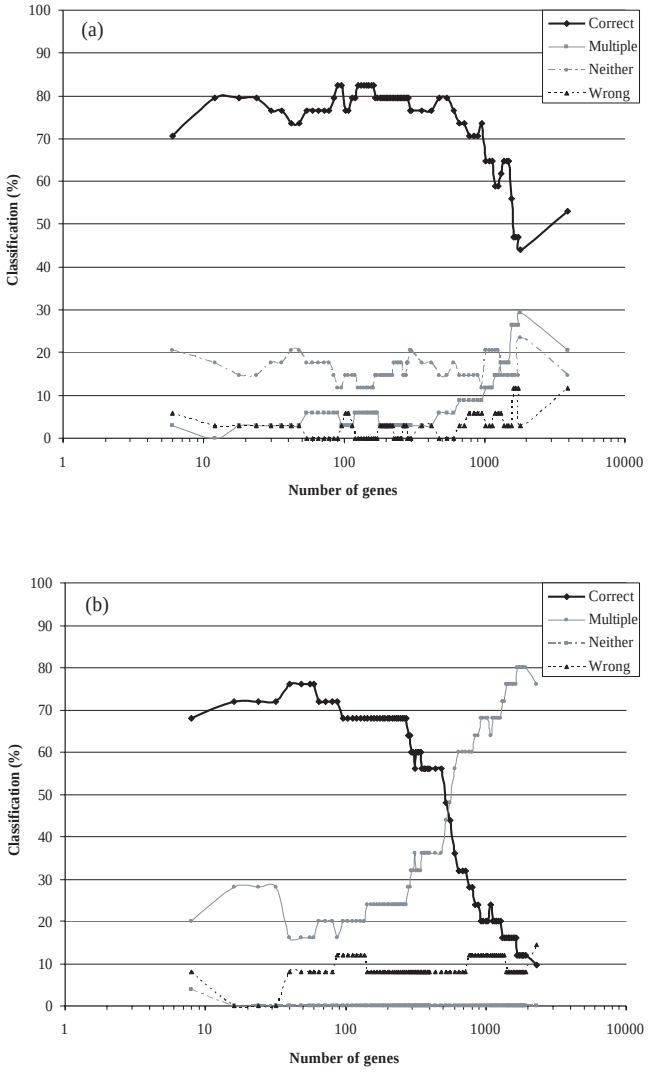

Fig. 1. Classification performance for different variable subsets. (a) Leukemia test set. (b) SRBCT test set.

Tables 10_SI, 11_SI, and 12_SI report the classification performances for different variable subsets. The best classification performance (76% of correctly classified samples) is obtained using the top 10 ÷ 15 variables of each class subset. In particular, 19 samples have been correctly classified, 2 samples have been misclassified and 4 recognized by more than one model. The misclassified samples are the two normal muscle tissues, Test 9 and Test 13 classified as RMS, while 3 out of 4 objects labeled as *multiple* are non-SRBCT cell lines (Test 3, Test 5, and Test 11). Comparing these findings with Khan's results, it can be noted that, similarly to SIMCA, the neural network committee scored equally Test 3, Test 5, and Test 11, while Test 9 and Test 13 received a higher vote for the

RMS class. Moreover, SIMCA best classification (76%) is comparable to the diagnostic performance of the neural network scheme (72%).

Table 13_SI lists the top 15 *class markers* maximizing the classifier performance and Figures 4 to 7 of **Supplementary information** show the expression profiles for some of them. The comparison of this subset with Khan's top ranked genes reveals that 41 of the 60 transcripts identified by SIMCA have been also selected by the neural network according to their relevance for the total classification. Finally, it is worthwhile comparing the number of features selected using the SIMCA scheme and determined by Khan's neural network. Indeed, SIMCA suggests using approximately 60 total variables while the neural network scheme achieved the best classification performance considering 96 genes. However, only 61 of the 96 markers selected by Khan are specific for a single class while the remaining 35 are over-expressed in more than one category. Since this condition induces a low variable ranking in the SIMCA-based feature extraction method, these latter 35 variables have not been identified as top-markers by the proposed procedure. Even if most of the genes specifically expressed in SRBCTs have not been previously related to the analyzed tumor types (Khan *et al.*, 2001), among the transcripts listed in Table 13_SI there are known EWS markers like *MIC2* and *GYG2*, genes reported to be expressed in RMS (e.g. *IGF2, MYL4, FGFR4, TNNT1*), in neuroblastoma cell lines (e.g. *MAP1B, MYO1B, NEF3, CRMP1*) and in Burkitt lymphomas (e.g. *CD10, HLA-DMA, ISG20, WAS*).

Reliability of the identified markers and classification results

To examine the reproducibility and stability of the selected genes, the marker identification and classification procedure has been repeated for different choices of the PCA parameters (e.g. number of selected components based on the total explained variance) and for varying compositions of the training set. Using the leukemia data set, four global models have been built automatically selecting the number of principal components that explained different amounts of the total variance (i.e. 50, 70, 80, and 90%). From the analysis of the top-30 genes identified by the four models (Table 14_SI), it can be inferred that the gene selection procedure is rather insensitive to the number of factors used to build the PCA models. Indeed, approximately 80% of the top-30 transcripts is conserved even in models described by varying number of principal components.

In addition to the training sets assigned by Golub *et al.* (1999) and Khan *et al.* (2001), SIMCA has been applied also to training and test sets obtained from random shuffles of the original assignments. Indeed, the reliability of feature selection and classification is a fundamental issue in model design, especially given the relatively small sample size associated with microarray data from cancer studies. When not enough samples are available for an extensive cross-validation and generalization analysis (*jackknife method*), a technique to assess the independence of selected features and classification performances from the chosen training set is to perform randomizations (*permutations*) of the original data set. Thus, $P = 100$ permuted data sets have been randomly generated starting from all leukemia and SRBCT samples. In particular, 100 equal random splits of 36 training and 36 test samples have been originated from the 72 leukemia specimen. Figure 2a shows the distribution of classifications (percentage of correctly classified samples) from the randomization analysis (Table 15_SI) as compared to the classification performance obtained using Golub's original train/test split (i.e. 82.3%, solid line). Both Figure 2a and Table 15_SI show the substantial stability of the estimates, given the small number of available samples and the inherent variability of the data. Similarly, Table 16_SI and Figure 8_SI show the percentage of the top-30 class-markers that are conserved when models are calibrated on random training sets as compared to those obtained using the original training set. Although the amount of non- random overlap is highly significant for ALL-B and ALL-T classes (i.e. on average 73 and 82%), the selection of AML-marking genes is more dependent on the populations used to calibrate the models (i.e. 51% of conserved markers). This result, considering also similar findings obtained by Li *et al.* (2001) using a genetic algorithm for gene selection, seems more a peculiarity of the leukemia data set than a general flaw of the method.

The same procedure has been applied to the SRBCT data set and Figure 2b shows the distribution of classifications from the randomization analysis (Table 17_SI). The classification performance obtained using Khan's original 63/25 partition (i.e. 76%) is reported for comparison (solid line) and confirms the stability of the estimates. Table 18_SI and Figure 9_SI show the percentage of the top-15 markers of each class that are conserved when the models are calibrated on random training sets as compared to those obtained using the original training set. In this second case study, the average percentage of conserved markers (i.e. 84, 79, 84, and 78% for EWS, RMS, NB, and BL respectively) clearly indicates that the selected set of genes is reliable in all the four classes and almost independent from the samples used to calibrate the models.

CONCLUSION

Although several computational schemes have been successfully applied to pair-wise analysis of tumor types from gene expression data, only a very limited number of marker selection methods and classification algorithms

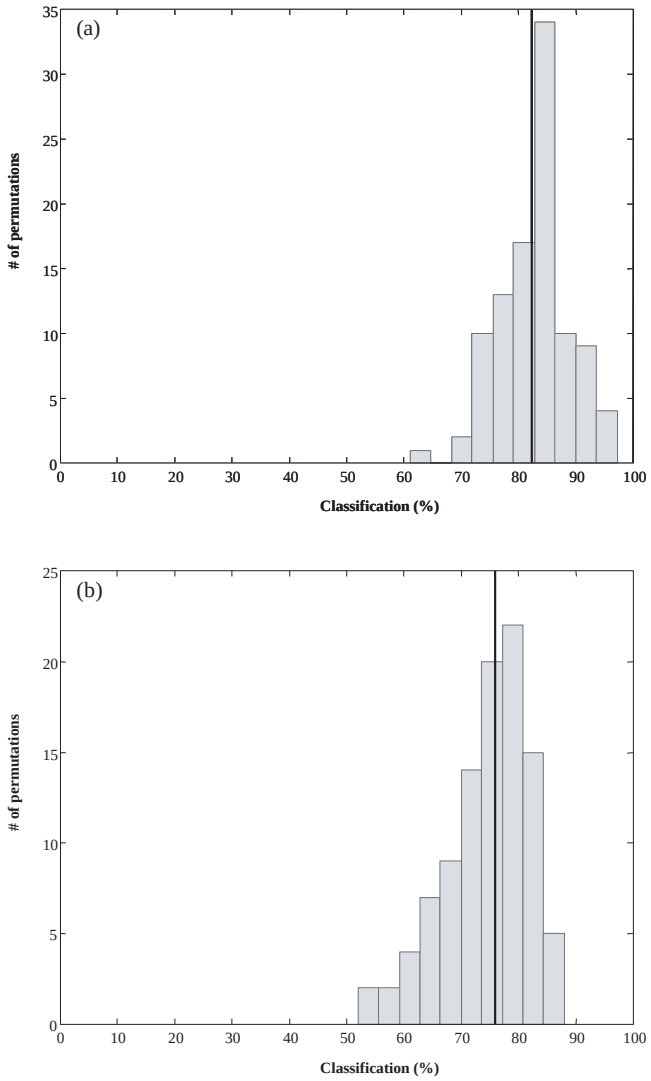

Fig. 2. Distribution of classifications (percentage of correctly classified samples) from randomization analysis. (a) Leukemia data set. In solid line is the classification performance obtained using the original train/test split (i.e. 82.3%). (b) SRBCT data set. In solid line is the classification performance obtained using the original train/test split (i.e. 76%).

have been developed in the context of multiple tumor categories. Specifically, Golub's group at MIT Whitehead Institute performed multiclass classification combining several binary classifiers (e.g. weighted voting, *k*-nearest neighbors, and support vector machines) in conjunction with different combination approaches (e.g. one-against-all, hierarchical partitioning; Yeang *et al.*, 2001; Ramaswamy *et al.*, 2001; Pomeroy *et al.*, 2002). Genes that correlate with each tumor class are identified by sorting all transcripts according to their signal-to-noise values (Golub *et al.*, 1999). Instead, Khan and co-workers applied artificial neural networks (ANN) to analyze cancer specimens belonging to different diagnostic categories. The ANN-based models have been trained to predict at the output layer the classification of previously unseen samples and the sensitivity of the classification has been linked to changes in gene expression levels thus identifying a subset of discriminating transcripts. The calibration of the model required the preliminary reduction of the feature space through principal component analysis since neural networks are prone to over-fitting if the analyzed system, as in the case of tumor expression profiling, is described by thousands of variables. Recently, Nguyen and Rocke (Nguyen and Rocke, 2002) extended the Partial Least Squares (PLS) procedure, previously presented for binary classification, to the analysis of cancer samples from multiple classes and Stephanopoulos *et al.* (2002) proposed a method based on Fisher discriminant analysis.

In this context, the present work addresses the implementation of a multivariate procedure that allows *marker identification* by extracting transcriptional features of physiological state and *sample diagnosis* by classifying a tumor specimen through the supervised analysis/comparison of expression profiles from multiple tumor types. The gene selection and sample classification scheme is based on Soft Independent Modeling of Class Analogy (SIMCA) and relies on the calibration of a principal component model for each class present in the analyzed data set (*disjoint class models*). In the context of gene expression analysis of multiple tumor types, the original SIMCA design has been adapted to solve the critical issue of feature selection. In particular, specific subsets of genes most highly correlated with several tumor categories have been identified examining the variance structure explained by each model, evaluating the performance of the classification scheme, and selecting those feature subsets that guarantees the maximal within-class modeling power as well as between- class separability. SIMCA procedure addresses the multiclass analysis directly with no need to design and combine binary classifiers or preliminarily reduce the feature space.

This multiclass modeling approach has been applied to two gene expression databases of different human tumor classes and the method has been able to identify groups of genes that could represent bases for subsequent experimental investigations. Moreover, the classification procedure has been able to distinguish with accuracy and robustness between multiple tumor subtypes.

Recently, the method has been also applied to a re-examination of the acute lymphoblastic leukemia database presented by Armstrong *et al.* (2002). This analysis allowed distinguishing samples carrying the t(4;11) chromosomal translocation, a *mixed-lineage leukemia gene (MLL)* genetic aberration characterized by frequent occurrence and adverse prognosis in infants.

Based on gene expression values, t(4;11) positive ALL's present a unique profile and can be separated from both ALL's negative for this translocation and MLL involving rearrangement with chromosome partners other than chromosome 4. In addition, a t(4;11)-specific set of markers has been identified and further confirmed by quantitative immunophenotyping flow cytometry (manuscript in preparation).

ACKNOWLEDGEMENTS

The authors are grateful to R.T. Kamimura for valuable comments and revision of the manuscript. G. te Kronnie and G. Basso at the Department of Pediatrics of the University of Padova are also acknowledged for helpful discussions on childhood malignancies.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Boldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Armstrong,S.A., Staunton,J.E., Silverman,L.B., Pieters,R., den Boer,M.L., Minden,M.D., Sallan,S.E., Lander,E.S., Golub,T.R. and Korsmeyer,S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.

Bittner,M., Meltzer,P., Chen,Y., Jiang,Y., Seftor,E., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Chow,M.L., Moler,E.J. and Mian,I.S. (2001) Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol Genomics*, **5**, 99–111.

De Maesschalck,R., Candolfi,A., Massart,D.L. and Heuerding,S. (1999) Decision criteria for soft independent modeling of class analogy applied to near infrared data. *Chemometr. Intell. Lab.*, **47**, 65–77.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

John,G., Kohavi,R. and Pfleger,K. (1994) Irrelevant features and the subset selection problem. In*: Machine Learning: Proceedings of the Eleventh International Conference (ICML-94)*. Morgan Kaufmann, San Fransisco, CA, pp. 121–129.

Joliffe,I.T. (1986) *Principal Component Analysis*. Springer, New York.

Keller,A.D., Schummer,M., Hood,L. and Ruzzo,W.L. (2000) Bayesian classification of DNA array expression data. *Technical report UW-CSE-2000-08-01*.

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.

Li,L., Winberg,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.

Massart,D.L., Vandeginste,B.G.M., Deming,S.N., Michotte,Y. and Kaufman,L. (1988) *Chemometrics: a textbook*. Elsevier, Amsterdam.

Nguyen,D.V. and Rocke,D.M. (2002) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.

Perou,C.M., Jeffrey,S.S., van de Rijn,M., Rees,C.A., Eisen,M.B., Ross,D.T., Pergamenschikov,A., Williams,C.F., Zhu,S.X., Lee,J.C. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.

Pomeroy,S.L., Tamayo,P., Gaasenbeek,M., Sturla,L.M., Angelo,M., McLaughlin,M.E., Kim,J.Y., Goumnerova,L.C., Black,P.M., Lau,C. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,C.H., Angelo,M., Ladd,C., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expressionsignatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Stephanopoulos,G., Hwang,D., Schmitt,W.A., Misra,J. and Stephanopoulos,G. (2002) Mapping physiological states from microarray expression measurements. *Bioinformatics*, **18**, 1054–1063.

Wold,S. (1976) Pattern recognition by means of disjoint principal components analysis. *Pattern Recogn.*, **8**, 127–139.

Yeang,C.H., Ramaswamy,S., Tamayo,P., Mukherjee,S., Rifkin,R.M., Angelo,M., Reich,M., Lander,E., Mesirov,J. and Golub,T. (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17** (Suppl. 1), S316–S322.