Classification of human malignancies by machine learning methods using DNA microarray gene expression data

Giorgio Valentini
D.I.S.I. Dipartimento di Informatica e Scienze dell' Informazione, Universita' di Genova
I.N.F.M. Istituto Nazionale di Fisica della Materia, Unita' di Genova
Genova - Italia

ABSTRACT: DNA hybridization microarrays supply information about gene expression through measurements of mRNA levels of large amounts of genes in a cell. The large amounts of data generated by DNA microarray technology need suitable methods for their processing and analysis. In this paper we apply supervised machine learning methods to the classification and recognition of human malignancies using DNA microarray gene expression data. We show that Multi-Layer Perceptrons and Support Vector Machines can correctly distinguish cancerous from normal tissues and Error Correcting Output Coding ensembles of learning machines can classify different types of lymphoma. Moreover our experimental results confirm the existence of distinct tumoral diseases inside the class of diffuse large B-cell lymphoma, offering also insights into the role of sets of coordinately expressed

INTRODUCTION

DNA microarrays [1,14] provide us with a large amount of information about gene expression, offering a wide picture of the functional status of a cell.

This information can be used to refine the traditional classification of human cancer based on morphological and clinical parameters; although the current taxonomy takes also into account molecular characteristics, in some cases it cannot detect functional differences among classes. In fact, actually, many patients receiving the same diagnosis could have distinct diseases, as reported by different clinical courses and treatment responses.

Information obtained by DNA microarray technology gives a snapshot of the overall functional status of a cell, offering new insights into potential different types of cancer, discriminated on molecular and functional basis.

The large amount of data produced by this powerful analytic technique can be processed through machine learning methods, using both unsupervised and supervised approaches.

In a typical unsupervised approach, expression patterns of several hundreds or thousands of genes are obtained both from cancerous and non cancerous tissues; then clustering algorithms [2] are used to group together similar expression patterns corresponding to different cells, in order to correctly separate cancerous from normal samples. By this approach we can also discover new functional classes not detected by traditional classifications of tumours, sometimes corresponding to different diseases with different clinical courses [3].

Anyway, unsupervised methods cannot always correctly separate classes. Supervised methods can overcome this problem, exploiting "a priori" biological and medical knowledge on the problem domain.

In this paper we show how to apply supervised learning methods for tasks related to cancer classification using DNA microarray gene expression data. We tackle the problem of recognizing cancerous and normal tissues, classifying different types of human malignancies and also identifying groups of genes related to a specific type of human cancer.

We have used data of a specialized DNA microarray, named "Lymphochip", developed at Stanford University School of Medicine [4], specifically designed to study lymphoid and carcinogenesis related genes.

These data are very challenging from a machine learning standpoint, considering that they are constituted by a small number of 4026-dimensional samples.

In our first task we distinguish cancerous from normal tissues using the overall information available. This dichotomic problem is tackled using Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP) and linear perceptrons (LP). SVM have been recently used in microarray gene expression analysis of data [5].

In our second task we try to directly classify different types of lymphoma (a multiclass problem) using MLPs and Parallel Non linear Dichotomizers (PND), i.e. ensembles of learning machines based on output coding decomposition of a multiclass problem [6].

Finally, we attempt to identify groups of genes specifically related to the expression of a tumour phenotype, exploiting "a priori" biological knowledge about sets of genes and information provided by clusters of coordinately expressed genes, i.e. "expression signatures"[3].

The paper is structured as follows. In the next section we summarize the basics about DNA microarray technology and machine learning methods for analyzing DNA microarray data. Then we present our experimental approach to the recognition of human lymphoma using gene expression data. We show three types of classification problems related to the analysis of DNA microarray data: classification of cancerous and non-cancerous lymphoid tissues, classification of different types of lymphoma and identification of diffuse large B-cell lymphoma subgroups. After a brief description of the supervised methods applied in our experimentation, the results are presented and discussed. Conclusions and future developments of this work end the paper.

DNA MICROARRAY FOR GENE EXPRESSION MONITORING

DNA hybridization microarrays [1] supply information about gene expression through measurements of mRNA levels of large amounts of genes in a cell.

Although regulation of protein amounts is not accomplished solely by regulation of mRNA, these methods offer a snapshot of the overall functional status of a cell: virtually all differences in cell type or state are related with changes in the mRNA levels of many genes.

DNA microarrays have been used in mutational analyses, genetic mapping studies and in genome monitoring of gene expression [1,2,5,9].

DNA microarray technology

DNA microarrays are microscopic arrays of DNA sequences printed on glass microscope slides. Using state of the art technologies all the human genome can be printed on a standard 1'' by 3'' microscopic slide in about one day.

The first step in preparing DNA microarrays involves the selection of suitable DNA targets. For organisms as mouse and humans, individual cDNA clones from cDNA libraries can be used as the source of gene-specific targets in the arrays. Then an arrayer robot prints the DNA samples through a cluster of specialized printing tips on a standard microscopic slide: each DNA sample is printed in a precise and known position on the slide.

The next step consists in the preparation of the fluorescent cDNA probes that will be used to hybridize the arrays. The mRNA of the cell whose gene expression has to be studied is isolated and purified. The prepared mRNA is used as template for synthesis of fluorescent cDNA probe by means of reverse transcription. Usually fluorecently labeled deoxyribonucleotides are used for producing cDNA probes. The cDNA sequences obtained by reverse transcription are then hybridized with the DNA samples printed on the microarray. A laser beam successively scans the slide and a raster image of the array is acquired. Measuring the fluorescent intensities of the image we can reconstruct the quantities of cDNA that hybridizes with each individual sample on the printed microscope slide. Consequently we gain information about the quantities of mRNA produced by the cell; i.e. we have a quantitative image of the gene expression.

Unfortunately the absolute representation of every RNA species in any cell or tissue sample cannot be obtained, as there is a complex relationship between the amount of input mRNA for a given gene and the intensity of the fluorescent cDNA probes, depending on a multitude of experimental conditions. Using relative representation of RNA species in two or more samples we can bypass these problems and moreover we are interested in differences in gene expression between samples, not in the absolute amounts of RNA.

For these reasons the ratios between two differently labeled cDNA probes, one of them acting as reference, are usually considered.

Gene expression data of different cells or different experimental/functional conditions are collected in matrices for numerical processing: each row corresponds to the gene expression data of a specific cDNA clone relative to all the examples, and each column corresponds to the expression data of all the cDNA clones relative to a specific cell sample. Typically thousands of genes are used and analyzed for each microarray sample.

Machine learning methods for analyzing microarray data

This powerful analytic technique requires machine learning methods for analyzing and extracting significant knowledge from these large amounts of gene expression data. Both unsupervised and supervised approaches can be used for analysing gene expression data.

In a typical unsupervised approach, expression patterns of several hundreds or thousands of genes are obtained for different cells or tissues or for different functional status of the same cell. Then clustering algorithms are used to group together similar expression patterns both for grouping sets of genes or sets of different cells (or different functional status of the same cell). By this approach we can discover genes functionally correlated [1] or we can separate expressions patterns of normal from pathological tissues [3].

Anyway, unsupervised methods cannot always correctly separate classes. Supervised methods can overcome this problem, exploiting "a priori" biological and medical knowledge on the problem domain. For instance, if we know from biological and medical knowledge that some genes are related to carcinogenesis, we can use the corresponding gene expression data for detecting tumoral cells, or we can learn to distinguish between normal and pathological tissues from known samples of cells, using in both cases supervised learning methods.

CLASSIFICATION OF LYMPHOMA USING DNA MICROARRAY DATA

In this paper we apply supervised machine learning methods for analyzing DNA microarray data with genes preferentially expressed in normal and cancerous lymphoid cells. Our aim consists in evaluating if supervised methods can be used to distinguish normal and cancerous lymphoid cells, and in assessing if supervised methods, exploiting "a priori" biological knowledge, can offer insights into the role of particular groups of genes in carcinogenic processes of lymphoid cells.

The data

In particular we have used data of a specialized DNA microarray, named "Lymphochip", developed at the Stanford University School of Medicine [4], specifically designed to study lymphoid and carcinogenesis related genes.
Data used in our experimentation[1] consist in 96 tissue samples from normal and cancerous populations of human lymphocytes, considering for each sample 4026 different genes preferentially expressed in lymphoid cells or with known roles in processes important in immunology or cancer.
We consider three main classes of lymphoma: diffuse large B-cell lymphoma (DLBCL), follicular lymphoma (FL) and chronic lymphocytic leukaemia (CLL) together with transformed cell lines (TCL) and normal lymphoid tissues [3].

Tab. 1 DNA microarray samples

Type of tissue	**Number of samples**
Normal lymphoid cells	24
DLBCL	46
FL	9
CLL	11
TCL	6

The total number of cancerous samples is 72, while the number of different non-cancerous samples amounts to 24.
Then we have a relative small amount of samples in a high dimensional (4026) space.
Gene expression data are expressed as fluorescence ratios normalized subtracting for each value the median between all the values. Missing gene expression data (about 6% of all the data) have been replaced with zeros.

The classification problems

Our first task consists in distinguish cancerous from normal tissues using the overall information available, i.e. all the 4026 gene expression data.
In our second task we try to directly classify different types of lymphoma (a multiclass problem), using again all the available gene expression data.

[1] The "Lymphochip" DNA microarray data are available at http://llmpp.nih.gov/lymphoma

The third task tries to validate the hyphothesis of Alizadeh et al. [3] about the existence of two distinct functional types of lymphoma inside DLBCL. Their clustering studies showed that two subgroups of DLBCL that they named germinal centre B-like DLBCL (GCB-like) and Activated B-like DLBCL (AB-like) can be separated using subsets of related genes. In our experimentation we attempt to identify groups of genes specifically related to the expression of these two different tumour phenotypes. In order to attain this goal, we employ "a priori" biological knowledge about sets of genes and information provided by clusters of coordinately expressed genes, i.e. "expression signatures" [3,9]. A gene expression signature is a set of genes identified by the biological process in which its component genes are known to function (i.e. proliferation), or by the cell type in which its component genes are expressed (i.e. germinal center B-cell). Using these subsets of genes to classify cancerous and normal tissues we can indirectly gain information on the significance of sets of correlated genes in carcinogenic processes.

Methods

For the first task (a dichotomic classification problem) we have used Support Vector Machines (SVM) [10], Multi Layer Perceptrons (MLP) and a linear perceptron. SVM are two-class classifiers theoretically founded on Vapnik's Statistical Learning Theory [11]. They act as linear classifiers in an high dimensional feature space originated by a projection of the original input space: the resulting classifier is in general non linear in the input space and it achieves good generalization performances maximizing the margin between the classes.

For the second multiclass problem we have used MLP, Parallel Linear Dichotomizers (PLD) and Parallel Non linear Dichotomizers (PND) [6].
PLD are ensembles of linear perceptrons, and PND are ensembles of MLP based on output coding decomposition of a multiclass problem [12]. These methods consist in decomposing a multi-class problem in a set of L of two-class problems according to some decomposition scheme, training the dichotomizers independently and combining the L outputs to give the class label. Learning machines based on decomposition methods are composed by two units: the Decomposition and the Decision Unit. In the decomposition unit a multiclass classification problem is decomposed in a set ot two class problems assigned to different dichotomic learning machines; each class is univocally determined by a codeword, i.e. a string of bits each of them learned by a different dichotomic classifier. The Decision unit reconstructs the original multiclass problem and selects the class through a suitable decoding function.

For the third problem we have used only SVM, as they showed good generalization performances in the first task

For the first two classification problems we used used 10-fold cross validation techniques for evaluating the generalization error of the learning machines together with the Joachims' $\xi\alpha$estimator [13] of the leave-one-out error for SVM. In the third classification problem we used only the $\xi\alpha$ estimator, as in the first classification task we achieved the same results about the estimation of the generalization error both using 10-fold cross validation and $\xi\alpha$ estimator of the leave-one-out error. In the two-classes classification problems we have estimated also the precision and recall [13]. The recall is the probability that a pattern of a positive class is correctly classified; the precision is the probability that a pattern classified as positive is indeed correctly classified.

In all learning tasks we have used *NEURObjects* [7], a set of C++ library classes for neural networks development[2] and *SVM*^{light} [8], a set of C applications implementing dichotomic SVM for classification tasks. We have also developed C++ software for preparing and analyzing input and output data.

Results

We have trained about 1500 SVM and 1500 MLP considering globally all the classification tasks involved in this experimentation. We have applied three different types of SVM, using linear, polynomial and radial basis kernel functions, and we have considered only MLP with one hidden layer both in two-class and multiclass MLP. The dichotomic base learner of the decomposition unit of the PND have been implemented by MLP with one hidden layer.
Gene expression data from the original data set have been transformed into a format suitable for SVM and MLP analysis.

Considering SVM, we have tried different values of the regularization parameter, varying it from 0.5 to 1000. We have used polynomials of degree from 2 to 5 and we have

[2] NEURObjects is on line available at
http://www.disi.unige.it/person/ValentiniG/NEURObjects

varied the sigma value of the radial basis SVM from 0.01 to 100.
We have used MLP with 1 hidden layer and we have performed model selection varying the number of hidden neurons from 3 to 15, and the parameters of the backpropagation algorithm.

Classifying cancerous versus non-cancerous tissues

In the first classification task we have used all the available gene expression information (input patterns 4026-dimensional) for separating cancerous from normal lymphoid cells. The results are shown in Tab. 2.

Linear SVM achieve the best results, but also MLP show an estimated generalization error of about 2% (using 10-fold cross validation). Interestingly the $\xi\alpha$ estimation of the leave-one-out error for SVM is identical to the computed estimation of the generalization error through 10-fold cross validation for all the 3 types of SVM. SVM show also a very high estimation (100%) of the probability of detecting tumoral lymphoid cells (recall), no matter the type of kernel function used.
Radial basis SVM show an high misclassification rate, due entirely to the low precision of this type of SVM (in fact the recall is 100%). This type of SVM has an high estimated Vapnik Chervonenkis (VC) dimension [11], confirmed also by the fact that sistematically all the input patterns are support vectors. Conversely linear SVM show a low estimated VC dimension and can correctly separate the two classes directly in the input space: it is likely that the local nature of the radial basis SVM in this case yields to overfitting, considering that we have also a small data set associated with a very high dimension of the input data.

Gloabally, we can state that supervised learning machines methods correctly separate cancerous from normal lymphoid tissues.

Classifying different types of lymphoma

In our second task we have classified directly different types of lymphoma, considering all the classes listed in Tab. 1. For this task we have used multiclass MLP, Parallel Linear Dichotomizers (PLD), One-Per-Class Parallel Non linear Dichotomizers (OPC-PND) and Error-Correcting-Output-Coding Parallel Non linear Dichotomizers (ECOC-PND) ensembles [7]. For ECOC-PND we have used 15-bit ECOC codes generated by exhaustive algorithms [12]. Fig.1 shows the results obtained varying the number of hidden units both for multiclass MLP and for the base learners of PND ensembles.

OPC and ECOC PND achieve the best results, with an estimated generalization error (through 10-fold cross validation) of about 5%, but also simple MLP achieve slightly worse but similar results. PLD fail on this task, achieving an high estimated error rate (about 23%), revealing that simple linear classifiers cannot be used for this task. Analysis of the confusion matrix for PND ensembles shows that the errors are due to false positives DLBCL (whereas they are normal lymphoid cells) and false positives TCL (whereas they are DLBCL) sistematically repeated in different PND ensembles. We need more studies to verify if these negative results depend on the limited accuracy of the classifiers, the unlucky partitions of data for cross validation, the too small size of the available samples or on the biological features of the analyzed samples.

Tab. 2 Classification of cancerous vs non-cancerous lymphoid cells: generalization error, precision and recall percent estimation through 10-fold cross validation. SVM-poly stands for polynomial SVM, SVM-rbf stands for radial basis SVM and LP stands for linear perceptron.

Learning machine model	**Gen. error**	**St. dev.**	**Precision**	**Recall**
SVM-linear	1.04	3.16	98.63	100.0
SVM-poly	4.17	5.46	94.74	100.0
SVM-rbf	25.00	4.48	75.00	100.0
MLP	2.08	4.45	98.61	98.61
LP	9.38	10.24	95.65	91.66

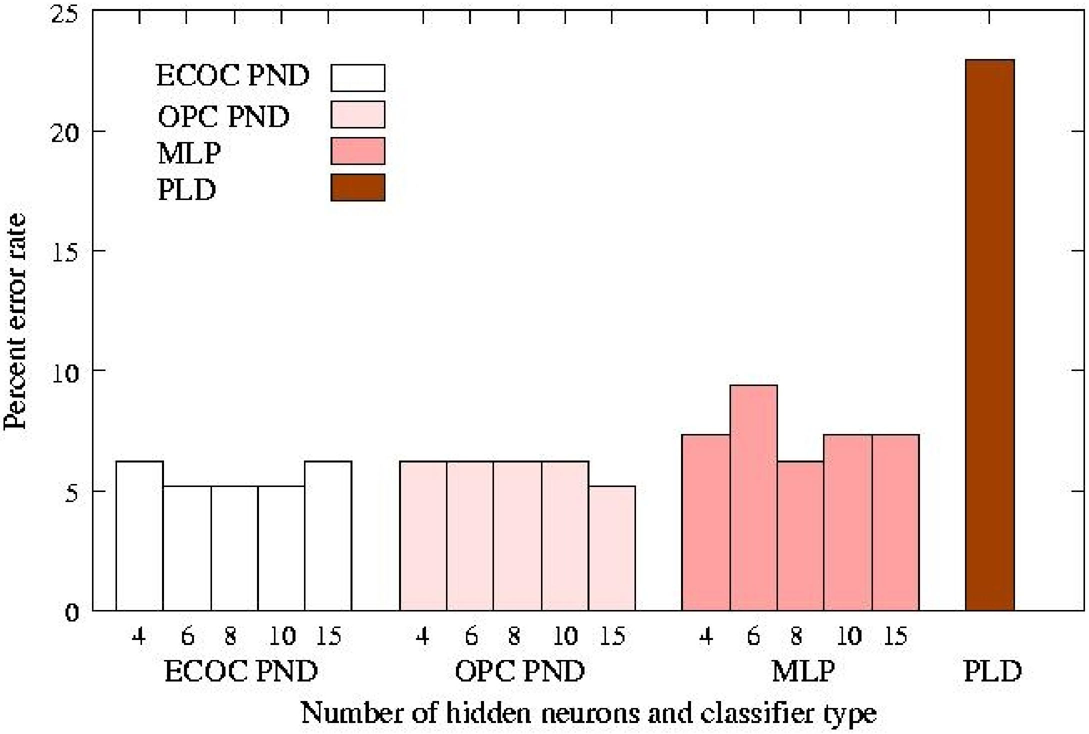

Fig. 1 Multiclass classification of different types of lymphoma using MLP multiclass, PLD and PND ensembles of learning machines.

Identifying DLBCL subgroups

Using clustering methods, Alizadeh et al. [3] showed that two subgroups of DLBCL lymphoma can be separated. They identified two subgroups of molecularly disitinct DLBCL: germinal centre B-like cells characterized by expression of genes normally expressed in germinal center B-cells and activated B-like cells characterized by expression of genes normally induced during in vitro activation of B cells. These two classes correspond also to patients with very different prognosis: those with activated B-like cells showed a significantly lower overall survival after treatment with comparable multi-agent chemoterapy regimens.

Lossos [9] and Alizadeh [3] showed that different subsets of genes could be responsible for the distinction of these two DLBCL subgroups: the expression signatures related to proliferation, T cell, lymphnode, and genes that distinguish

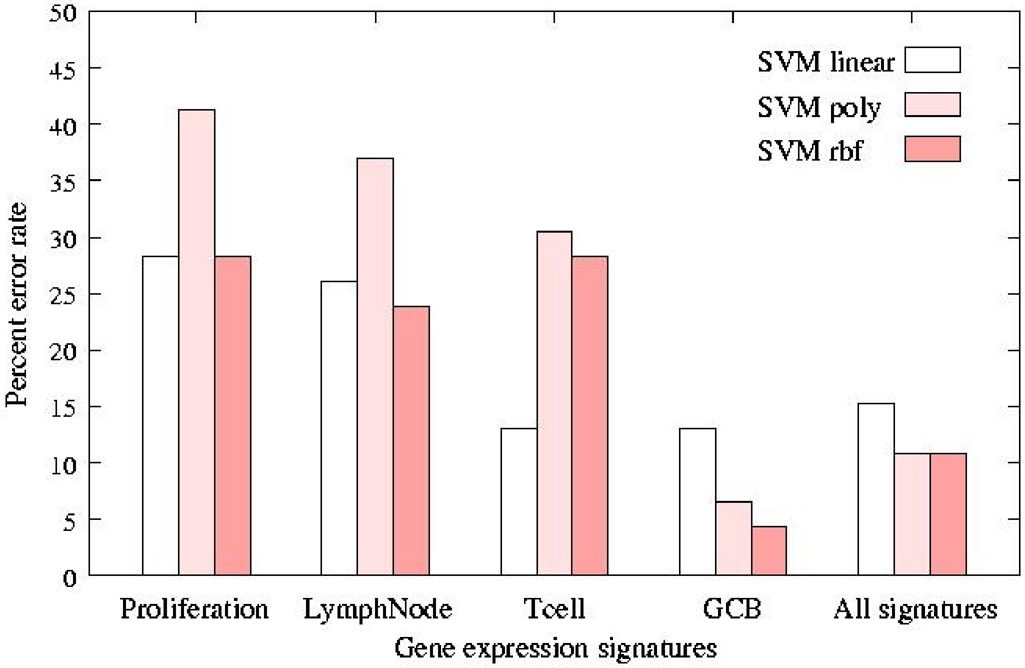

Fig. 2 Estimated generalization error for the classification of GCB-like and AB-like subgroups of DLBCL using 4 different gene expression signatures (and the 4 signatures alltogether). SVM poly stand for polynomial SVM and SVM rbf stands for radial basis SVM.

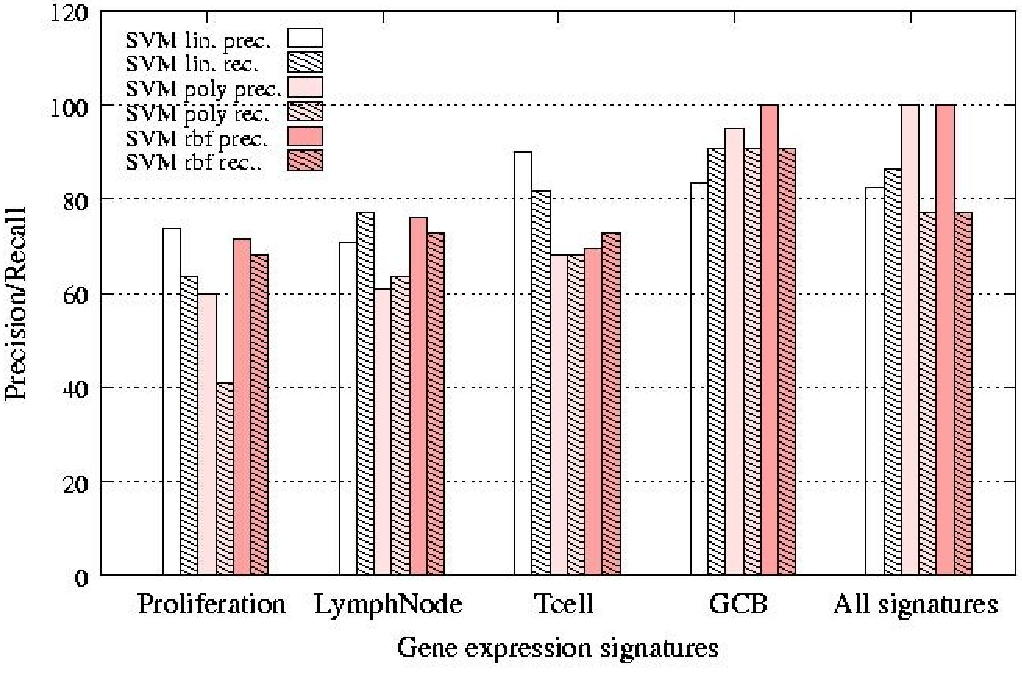

Fig.3 Estimated precision and recall for the classification of GCB-like and AB-like subgroups of DLBCL using 4 different gene expression signatures and the 4 signatures alltogether.

germinal centre B-cells from other stages in B-cell ontogeny (GCB expression signature) showed differential gene expressions between these two subgroups.

In our experimentation we have employed "a priori" biological knowledge about sets of genes and information provided by clusters of coordinately expressed genes (expression signatures) in order to verify if we can identify an expression signature related to a DLBCL partition proposed by Alizadeh.

More preisely, we have performed 5 classification tasks, using SVM and leave-one-out methods for estimating the generalization error on the dichotomic classification task for separating germinal center B-cells and activated B-like cells. For each classification task we have used a different expression signature from the four listed above, and we have also tried the same classificaction task using the above signatures alltogether. The results are shown in Fig. 2 and Fig. 3.

Using the "Proliferation" and the "Lymphnode" signature we cannot separate these two subgroups and also using only the Tcell signature we can separate the two subclasses only with an estimated error of about 13%, a precision of about 90%and an estimated recall of about 81% using linear SVM. The best results are achieved with the GCB expression signature, with an estimated generalization error of about 4% and an estimated precision of 100% (estimated recall 91%) using radial basis SVM. Using all the signatures we obtain an high precision (100%) both with polynomial and radial basis SVM but a generalization error of about 10%.

These results confirm the hyphotesis of Alizadeh about the existence of two distinct subgroups in DLBCL and identify the GCB signature as a cluster of coordinately expressed genes highly related to the separation between the GCB-like and AB-like DLBCL subgroups.

CONCLUSIONS

This paper shows that supervised machine learning methods can be applied to the analysis of DNA microarray gene expression data.

We have performed three classification tasks for the analysis of gene expression data related to diffuse large B-cell lymphoma.

In the first two tasks we have shown that SVM, MLP and PND can be successfully applied to the classification of cancerous and normal lymphoid tissues and to the recognition of different types of lymphoma.

In the third task we pointed out how to use "a priori" biological and medical knowledge for separating two functional subclasses of DLBCL not detectable with traditional morphological classification of lymphoma, identifying a set of coordinately expressed genes related to the separation of the two DLBCL subgroups.

The analysis of the causes of the classification errors of the SVM on the third task remains an open problem: we need more experimentation (and probably more data) for establishing if the SVM errors are due to the errors of the clustering algorithm [2] that have defined the two DLBCL subclasses or to the limited accuracy of SVM or to the too small number of the available samples.

Developments of this work could consist in integrating "a priori" biological knowledge, supervised machine learning methods and unsupervised clustering methods for discovering distinct subclasses of malignancies based on functional and molecular differences. These tasks are fundamental to detect types of tumours not detectable with traditional methods, enabling physicians to select therapies targeted to specific tumoural diseases.

ACKNOWLEDGMENTS

This work was partially funded by I.N.F.M. Unità di Genova and D.I.S.I. – Università di Genova.

REFERENCES

1. J. De Risi et al. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680-686, (1997).
2. M.B. Eisen et al. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95, **25**, 14863-14868, (1998).
3. A. Alizadeh et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403, 503-511, (2000).
4. A. Alizadeh et al. The Lymphochip: a specialized cDNA microarray for genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harbor Symp. Quant. Biol.* (in press).
5. M. Brown et al. Knowledge-base analysis of microarray gene expression data by using Support Vector Machines. *PNAS*, 97, **1** , 262-267 (2000).

6. F. Masulli and G. Valentini. Parallel Non linear Dichotomizers. IJCNN2000, *The IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 2, 29-33 (2000).
7. G. Valentini and F. Masulli. NEURObjects: a set of library classes for neural network development. *IIA'99-SOCO'99*, ICSC Academic Press, Millet, Canada, 184-190 (1999).
8. T. Joachims. Making large-scale SVM learning practical. *Advances in Kernel Methods – Support Vector Learning*, B.Scholkopf, C. Burges and A. Smola (ed.) MIT Press (1999).
9. I. Lossos et al. Ongoing immunoglobulin somatic mutation in germinal center Bcell-like but not in activated B cell-like diffuse large cell lymphomas. *PNAS*, 97, **18**, 10209-10213 (2000).
10. C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, 20, **3**, 273-297 (1995).
11. V. Vapnik. *Statistical Learning Theory*. Wiley & Sons, New York (1998).
12. T. G. Dietterich and G. Bakiri, Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research* 2, 263-286 (1995).
13. T. Joachims, Estimating the Generalization Performance of a SVM Efficiently. *Proceedings of the International Conference on Machine Learning*, Morgan Kaufman (2000).
14. D.J. Lockhart and E.A. Winzeler. Genomics, gene expression and DNA arrays. *Nature*, 405, 827-836 (2000).