Marker Identification and Classification of Cancer Types Using Gene Expression Data and SIMCA

S. Bicciato, A. Luchini, C. Di Bello
Department of Chemical Process Engineering, University of Padova, Italy

Summary

Objectives: High-throughput technologies 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 DNA microarrays to monitor expression profiles from tumor cells is improving cancer analysis to levels that classical methods have been unable to reach. However, molecular diagnostics based on expression profiling requires addressing computational issues as the overwhelming number of variables and the complex, multiclass nature of tumor samples. Thus, the objective of the present research has been the development of a computational procedure for feature extraction and classification of gene expression data.

Methods: The Soft Independent Modeling of Class Analogy (SIMCA) approach has been implemented in a data mining scheme, which allows the identification of those genes that are most likely to confer robust and accurate classification of samples from multiple tumor types.

Results: The proposed method has been tested on two different microarray data sets, namely Golub's analysis of acute human leukemia [1] and the small round blue cell tumors study presented by Khan et al. [2]. The identified features represent a rational and dimensionally reduced base for understanding the biology of diseases, defining targets of therapeutic intervention, and developing diagnostic tools for classification of pathological states.

Conclusions: The analysis of the SIMCA model residuals allows the identification of specific phenotype markers. At the same time, the class analogy approach provides the assignment to multiple classes, such as different pathological conditions or tissue samples, for previously unseen instances.

Keywords

Gene expression data, SIMCA, PCA, feature extraction, classification

Methods Inf Med 2004; 43: 4–8

1. Introduction

High-throughput technologies are radically boosting the understanding of living systems, thus creating enormous opportunities to identify target genes and pathways for drug development and to elucidate networks of genomic regulation by the comparison of the phenotype of cells in different physiological states. At present, in the so-called *post-genomic era,* the accent in biological research is shifting from data acquisition to data analysis and interpretation. Indeed, the increasing pace of genomic data accumulation poses the challenge to develop analysis procedures able to generate new knowledge and upgrade the information content of these databases.

Several different methods have been proposed to analyze large amounts of expression profiling data and identify set of genes that can serve as diagnostic platforms. Among all, the most widely used technique is hierarchical agglomerative clustering. As reported in many publications, clustering techniques have been applied to identify groups of genes sharing similar expression profiles and the results obtained so far are extremely valuable. Clustering techniques have been demonstrated to be useful tools in grouping functional related families of genes. However, clustering methodologies represent an example of unsupervised analysis that is not appropriate for the incorporation of prior knowledge about the observations, as for example sample labels (i.e., normal or tumor tissue), in the partitioning and grouping procedure. As such, cluster analysis may not be a good framework for diagnosis or classification of diseases nor to pinpoint specific features marking a phenotype.

Several machine learning methods have been applied to classify pathologies and tissue samples on the basis of their expression profiles [1, 2]. This task presents a major challenge due to the overwhelming number of variables (genes), the majority of which is not relevant to the description of the problem and could potentially degrade the performance of the classification scheme by masking the contribution of the relevant features. Thus, together with the development of classification schemes, it is of paramount importance to identify those genes that are most likely to confer high classification accuracy *(gene selection)*. Indeed, these key informative features represent a base of reduced cardinality for subsequent experimental investigation aimed at determining their role, if any, in the generation and progression of the analyzed phenotype.

The purpose of this work is to present a procedure for detecting patterns of expression correlated to peculiar phenotypes through a supervised analysis of labeled samples in the context of multiple tumor types. Specifically, it presents results from a computational framework based on principal component analysis (PCA) and on the Soft Independent Modeling of Class Analogy (3). This approach simultaneously allows identifying specific markers of phenotypes and predicting the class label of a set of previously unseen instances.

The properties of principal component analysis are used to implement a modeling scheme called SIMCA, which has been previously applied to solve many pattern recognition and classification problems. In a multi-class problem, SIMCA works considering each class separately. For each class, a principal component analysis is performed leading to a different PCA model for each class (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 unknown's data and each class model. The model that best fits the unknown, if any, represents the class assigned to that sample. Even if reliable classification of previously unseen instances is the ultimate goal of this approach, SIMCA can also be used for the fundamental issue of feature selection. Indeed, examining the variance structure 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 tumor type distinctions.

The SIMCA modeling approach has been applied to the analysis of two gene expression databases, namely the data set from Golub's work on leukemia classification [1] and the study presented by Khan et al. [2] on small round blue cell tumors.

2. Materials and Methods

2.1 Gene Expression Data from Tumor Samples

Two gene expression data sets have been used to illustrate the gene selection and classification method. The leukemia study provides measurements for 3930 probes in 72 samples collected from acute leukemia patients. Forty-seven cases were diagnosed as acute lymphoblastic leukemia (ALL) and the other 25, as acute myeloid leukemia (AML). The ALL class was further subdivided in 38 B-lineage and 9 T-lineage ALL samples. RNA prepared from bone marrow and peripheral blood cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix (Santa Clara, CA). The second database consists of gene-expression data from cDNA experiments describing four childhood malignancies: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (EWS). The 63 training samples included both tumor biopsy material (13 EWS and 10 RMS) and cell lines (10 EWS, 10 RMS, 12 NB and 8 BL). An independent set of 20 blind test samples has been used for testing the classification capabilities of the proposed approach.

2.2 The SIMCA Method

Soft Independent Modelling of Class Analogy (SIMCA) uses PCA to extract different characteristics from a set of objects. These characteristics are then used to divide the set into different classes, defined by the user. An object is classified as belonging to the class it resembles the most. The method is discussed in [3-4].

PCA is a statistical data analysis technique that allows reducing the dimensionality of the system while preserving information on variable interactions [4]. PCA transforms the original variables into a set of linear combinations, the principal components (PC), with special properties in terms of variances. Specifically, it determines an optimal linear transformation $\mathbf{y} = \mathbf{Wx}$ of an *n-dimensional* data vector $\mathbf{x}$ into another *m-dimensional* ($m \leq n$) transformed vector $\mathbf{y}$. The *mxn* fixed linear transformation matrix $\mathbf{W}$ is designed exploring statistical correlations among the variables of the original data matrix and finding reduced compact data representations that retain maximum nonredundant and uncorrelated intrinsic information of the original data. Exploration of the original data set is based on computing and analyzing the data covariance matrix, its eigenvalues and corresponding eigenvectors organized in descending order. Each element of the *m-dimensional* transformed feature vector $\mathbf{y}$ will be linearly independent and in decreasing order according to decreasing information content. This allows a straightforward reduction of the dimensionality by discarding the feature elements with lower information content. Thus, all original *n-dimensional* data patterns can be optimally transformed to data

Table 1 B-ALL top 20 markers

Accession number	Symbol	Name
X82240	TCL1A	T-cell leukemia/lymphoma 1A
M89957	CD79B	CD79B antigen (immunoglobulin-associated beta)
L33930	CD24	CD24 antigen (small cell lung carcinoma cluster 4 antigen)
L08895	MEF2C	MADS box transcription enhancer factor 2, polypeptide C
Z49194	POU2AF1	POU domain, class 2, associating factor 1
U52682	IRF4	Interferon regulatory factor 4
X55740	NT5E	5' nucleotidase, ecto (CD73)
D88270	VPREB1	(lambda) DNA for immunoglobin light chain
U05259	CD79A	MB-1 gene
U36922	FKHR	Fork head domain protein (FKHR) mRNA, 3' end
U18259	MHC2TA	MHC class II transactivator
K01911	NPY	Neuropeptide Y
U46006	CSRP2	Cysteine and glycine-rich protein 2
X99920	S100A13	S100 calcium binding protein A13
U10485	LRMP	Lymphoid-restricted membrane protein
M38690	CD9	CD9 antigen (p24)
M84371	CD19	CD19 gene
U49020	MEF2A	Myocyte-specific enhancer factor 2A, C9 form
X58529	IGHM	Immunoglobulin heavy constant mu
X74301	MHC2TA	MHC class II transactivator

patterns in a feature space with lower dimensionality. The algorithm chosen for this work is based on *singular value decomposition (SVD)* and, since several texts cover the calculation of the PC's in details (e.g., [4]), theoretical aspects will be omitted.

Once a principal component model is calculated, new object data can be projected onto the PC vector space and the total residuals between the predicted and the original data represent a measure of how well the projected data fit the original model. Comparing, through an F-test, the prediction errors with the residual limits calculated for the data used in the model construction (training set), it is possible to see if the sample belongs to the modeled class or not.

With the SIMCA technique, different classes are modeled individually by a separate principal component model. The number of significant PC's is determined for each class. The residuals are used for the creation of boundaries around each class. The distance or standard deviation s_k^K of object k, described by m variables, to a class K, modeled with p_K principal components, is given by the sum over the m variables of the distances (or residuals) between object k and the PC model along each variable j (e_{kj}^K).

A fundamental issue that can be addressed using the SIMCA modeling scheme is variable selection, meaning the identification of those peculiar features that better characterize a category. 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, iii) computation of the minimum number of variables needed to maximize multiclass classification. Specifically, comparing the different values of e_{kj}^K (i.e., the variance of variable j of object k in class model K), it is possible 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-K-variable* is defined so that it presents large values of the residuals when class-K-samples are fitted to all categories but the true K model and, at the same time, the error of the K model is minimized only by class-K-samples. The entire procedure has been implemented in Matlab.

3. Results and Discussion

3.1 Leukemia Data Set

Following the experimental setup described in [1], the data has been split into a training set consisting 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). With the aim to first quantify the relative relevance of each transcript in describing the three different subtypes of leukemia, three PCA models are built using the three groups of training samples after autoscaling the expression levels. 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 SIMCA models. The number of principal components has been determined using a leave-one-out cross-validation procedure (details described in [4]). Each of the 7129 variables have been assigned to one of the classes analyzing the sum of the residuals produced when a class-K-sample is fitted to all models but the true one, checking, at the same time, the unique minimization of model K residuals for class-K-samples only. The selected genes have been finally sorted combining their modeling power with their discriminating power, among the different classes. This procedure identifies 615 B-lineage ALL, 2657 T-lineage ALL, and 658 AML related transcripts. Tables 1, 2, and 3 list the top 20 markers for each of the three subtypes of leukemia. For all of these features, experimental evidences prove or suggest an important role in acute B-cell

Table 2 T-ALL top 20 markers

Accession number	Symbol	Name
X00437	TRB	T cell receptor beta locus
M28826	CD1B	CD1B antigen, b polypeptide
X76223	MAL	MAL gene exon 4
X04145	CD3G	CD3G antigen, gamma polypeptide (TiT3 complex)
X03934	CD3D	CD3D antigen, delta polypeptide (TiT3 complex)
U23852	LCK	T-lymphocyte specific protein tyrosine kinase p56lck (LCK)
X59871	TCF7	Transcription factor 7 (T-cell specific, HMG-box)
HG4128-HT4398		Anion Exchanger 3, Cardiac Isoform
U50743	FXYD2	FXYD domain-containing ion transport regulator 2
X14975	CD1E	CD1E antigen, e polypeptide
J04132	CD3Z	CD3Z antigen, zeta polypeptide (TiT3 complex)
M26692	LCK	T-lymphocyte-specific protein tyrosine kinase (LCK)
U40271	PTK7	PTK7 protein tyrosine kinase 7
U49835	CHI3L2	Chitinase 3-like 2
X87241	FAT	FAT tumor suppressor homolog 1 (Drosophila)
L10373	TM4SF2	Transmembrane 4 superfamily member 2
M23323	CD3E	T-cell surface glycoprotein CD3 epsilon-chain
U14603	PTP4A2	Protein tyrosine phosphatase type IVA, member 2
M37271	CD7	CD7 antigen (p41)
X60992	CD6	CD6 antigen

lymphoblastic, T-cell lymphoblastic, and myeloid leukemias, respectively [1].

The accuracy in the classification of the blind test set improves when using a subset of the modeling features, as compared to all the transcripts or to any random selection of them. Indeed, the classification accuracy arises from 53 % of correct predictions when using all the expression profiles to 82 % of correct predictions using the top 20÷40 markers identified by the SIMCA approach. The definition of unified criteria for the selection of an optimal (or near optimal) subset of markers is under development.

3.2 Small, Round Blue Cell Tumor Data Set

The 63 training samples included both tumor biopsy material and cell lines for a total of 4 different categories (EWS, RMS, NB, and BL). An independent set of 20 blind test samples has been used for testing the classification capabilities of the proposed approach. Similarly to case study 1, SIMCA modeling scheme identifies 600, 496, 512, and 700 genes related to EWS, RMS, NB, and BL respectively. For sake of space, Table 4 lists only the top 15 markers of each category. Most of these transcripts are included in Khan's list of top ranking genes [2]. The classification accuracy improved from 10 % of correct calls, obtained designing the classifier with all the 2308 genes, to 95 % when using only the top 10÷15 markers of each category.

4. Conclusions

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 is improving cancer analysis to levels that classical methods have been unable to reach. However, cancer analysis and classification on the basis of microarray data

Table 3 AML top 20 markers

Accession number	Symbol	Name
M84526	DF	D component of complement (adipsin)
Y00339	CA2	Carbonic anhydrase II
M27891	CST3	Cystatin C (amyloid angiopathy and cerebral hemorrhage)
M96326	AZU	Azurocidin gene
M20203	ELA2	Neutrophil elastase gene, exon 5
M31551	PAI2	Plasminogen activator inhibitor, type II (arginine-serpin)
M31166	PTX3	Pentaxin-related gene, rapidly induced by IL-1 beta
M27783	ELA2	Elastase 2, neutrophil
X95735	ZYX	Zyxin
U05572	MAN2B1	Mannosidase, alpha, class 2B, member 1
M30703	AR	Amphiregulin (AR) gene
M57731	GRO2	GRO2 oncogene
L08177	EBI2	Epstein-Barr virus induced gene 2
M21119	LYZ	Lysozyme (renal amyloidosis)
X97748	PTX3	PTX3 gene promotor region
J04990	CTSG	Cathepsin G
M23197	CD33	CD33 antigen (gp67)
U46751	SQSTM1	Sequestosome 1
M28130	IL8	Interleukin 8 gene
L05424	CD44	CD44 gene (cell surface glycoprotein CD44)

Table 4 Small, round blue cell tumors top 15 markers

EWS		RMS		NB		BL	
Clone ID	Symbol	Clone ID	Symbol	Clone ID	Symbol	Clone ID	Symbol
866702	PTPN13	244618		44563	GAP43	183337	HLA-DMA
770394	FCGRT	298062	TNNT2	135688	GATA2	767183	HCLS1
377461	CAV1	784224	FGFR4	395708	DPYSL4	740604	ISG20
43733	GYG2	461425	MYL4	812105	AF1Q	769657	PPP1R2
357031	TNFAIP6	770059	HSPG2	383188	RCV1	241412	ELF1
814260	FVT1	25725	FDFT1	629896	MAP1B	200814	MME
1473131	TLE2	789253	PSEN2	308231	MYO1B	80109	HLA-DQA1
1435862	MIC2	769716	NF2	377048	MYO1B	236282	WAS
52076	OLFM1	796258	SGCA	220096	CNGB1	624360	PSMB8
80338	SELENBP1	1409509	TNNT1	743229	NEF3	530185	CD83
1471841	ATP1A1	245330	IGF2	784257	KIF3C	609663	PRKAR2B
365826	GAS1	813841	PLAT	878280	CRMP1	47475	PIR121
308497	HT036	898219	MEST	325182	CDH2	840942	HLA-DPB1
364934	DAPK1	246035	--	823886	--	68977	PSMB10
767345	--	755750	NME2	842918	FARP1	814526	RNPC1

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).

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 tumor specimens 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. In the context of gene expression analysis, 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 and evaluating the performance of the classification scheme. SIMCA procedure addresses the multiclass analysis directly with no need to design and combine binary classifiers or preliminary reduce the feature space.

Proof of concept has been given through the analysis of two gene expression databases, namely the data set from a work on leukemia subtypes and a study on small round blue cell tumors. 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.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh M, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; 286 (5439): 531-7.
2. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001; 7 (6): 673-9.
3. Wold S, Sjostrom M. SIMCA: a method for analyzing chemical data in terms of similarity and analogy. Kowalski BR, editor. Chemometrics: Theory and Application. Washington: ACS; 1977.
4. Jolliffe I. Principal Components Analysis. Springer-Verlag; 1986.

Correspondence to:
Dr. Silvio Bicciato
Department of Chemical Process Engineering
University of Padova
Via Marzolo, 9
35131 Padova, Italy
E-mail: silvio.bicciato@unipd.it