EXTRACTING BIOMEDICAL INFORMATION FROM GENE EXPRESSION MICROARRAY DATA BY MULTIVARIATE CURVE RESOLUTION

Joaquim Jaumot [1], Romà Tauler [2] and Raimundo Gargallo [1]

1. Departament de Química Analítica, Universitat de Barcelona
2. Departament de Química Ambiental, I.I.Q.A.B. - C.S.I.C.

Diagonal 647, E-08028, Barcelona, Spain **E-mail: joaquim@apolo.qui.ub.es** **www.ub.es/gesq/eq1_eng.htm**

INTRODUCTION

DNA microarray technology has made possible to monitor gene expression levels for thousands of genes in a single experiment. From these experiments, information about the existence of patterns and relationships between samples (cell lines) and variables (genes) can be obtained. Because of the huge amount of data generated in a single experiment, data compression and data analysis methods are needed to extract and understand the information contained in the data.

Principal Components Analysis (PCA) and classification methods have been proposed as standard tools [1]. These methods allow the classification of several samples in an appropriate number of groups providing information about the relationships between the cell lines. Additionally, information about over- or under- expressed genes related to the apparition of cancer diseases can be obtained.

In this work, microarray data are proposed to be studied by using different clustering methods and by Multivariate Curve Resolution Alternating Least Squares (MCR-ALS).

DATA UNDER STUDY

Data set used in this work has been generated by Ross [2] and are publicly available at the web site http://discover.nci.nih.gov/nature2000/data/selected_data. mRNA was extracted from 60 cancer cell lines and hybridized to cDNA microarrays.

Type of cancer	Abbeviation	Nº of samples
Breast carcinoma	BR	8
Central nervous system tumor	CNS	6
Colon carcinoma	CO	7
Non-small cell lung cancer	LC	9
Leukemia	LE	6
Melanoma	ME	8
Ovarian Carcinoma	OV	6
Prostate carcinoma	PR	2
Renal carcinoma	RE	8

The analysis is based on 1416 variables corresponding to the 1375 genes showing the highest variance with some repetitions. The data are expressed as the log ratio between the gene expression level of each cell line and a reference made by mixture of 12 of the 60 cell lines. These data have been previously analyzed by PCA methods [3].

METHODS OF CLASSIFICATION

The goal of gene expression data clustering is to group together genes or samples that have similar expression profiles.

CLUSTERING

There are two approaches for clustering [4]:

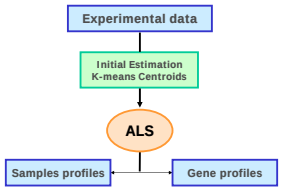
HIERARCHICAL clustering algorithms typically build a tree (dendogram) that represents the similarity between the different samples evaluated using a "distance" measure. Several linkage methods (centroids, Ward's method, nearest neighbour, ...) with different forms of measure distances (Euclidean, Pearson Correlation, Manhattan, ...) have been tested.

NON-HIERARCHICAL clustering algorithms are based on the selection of an arbitrary number of clusters. The members belonging to each cluster will be checked by distance and relocated into the more appropriate clusters with higher separability. One of the most common methods for non-hierarchical clustering is the k-means method that allows to obtain either the grouping of the samples and the variables.

PRINCIPAL COMPONENT ANALYSIS. Principal Component Analysis (PCA) [4] is another standard method for the analysis of microarray data. This method finds new axes explaining maximum variance. It allows the compression of the original information. Projection on the new PCA axes provides information about how samples are clustered.

MULTIVARIATE CURVE RESOLUTION is a factor analysis method that decomposes original data into few contributions of the components present in the system [5]. An Alternating Least Squares (ALS) algorithm is used for this bilinear data matrix decomposition.

In the case of microarray data analysis information about each cluster of samples (scores) and its associated gene profile (loadings) can be obtained. The centers of the clusters found by the k-means clustering method are proposed as initial estimations for the ALS optimization. In this work, application of constraints (like non-negativity for the sample profiles) can provide results with an easier interpretation of the biological signification.

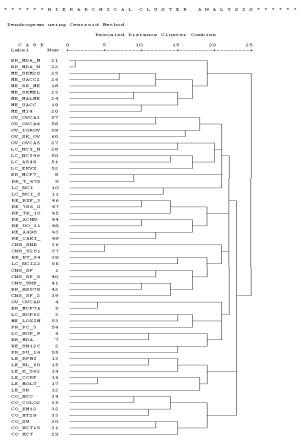


CLUSTERING

HIERARCHICAL

Several linkage methods with different ways of measuring distances have been tested.

Best linkage results were obtained by the centroid method using Pearson Correlation coefficient distances.



Both classification methods separate clearly leukemia, melanoma and colon carcinoma samples.

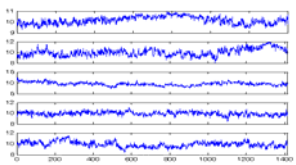
NON-HIERARCHICAL

Non-hierarchical classification has been performed using the k-means method.

Best results were obtained using 5 clusters:

CLUSTER	Nr. of Samples	Cluster Membership
1	25	7 RE, 6 CNS, 5 LC, 3 OV, 3 BR, 1 ME
2	9	7 ME, 2 BR
3	6	6 LE
4	8	3 LC, 2 PR, 1 OV, 1 BR, 1 RE
5	11	7 CO, 2 OV, 1 BR, 1 LC

The center of the final clusters for each of the 1416 gens were also obtained using k-means linkage method.

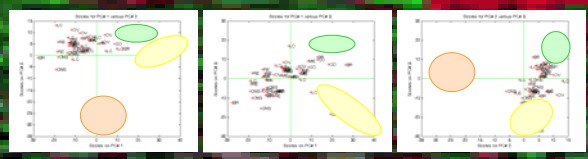


These genic profiles have been used as an initial estimation for the MCR-ALS method.

PRINCIPAL COMPONENT ANALYSIS

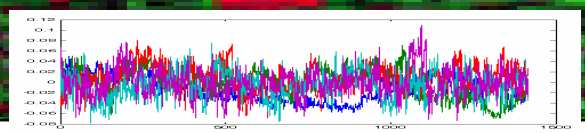
Analysis of SCORES

Clusters of samples can be observed in the plots of the first 3 PCA scores. Groups of ME, LE & CO are clearly identified.



Analysis of LOADINGS

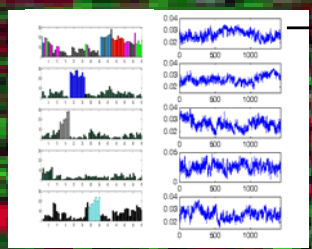
Correlation among different genes can be deduced from the analysis of loading profiles.



MULTIVARIATE CURVE RESOLUTION

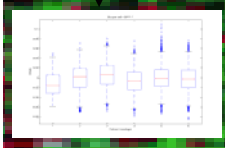
MCR-ALS results show that each MCR-ALS component is characterized by:

MCR-ALS Component	Type of cancer
1	CNS, OV, RE, LC
2	ME
3	LE
4	Not well defined
5	CO



Information about samples

Information about gene profiles



Identification of outliers in boxplots of resolved gene profiles allow an easy determination of genes under or over- expressed.

CONCLUSIONS

Extraction of information from DNA microarray data can be performed using several chemometrics methods. A comparison of results obtained in this work shows that different methods provide similar results.

Application of Multivariate Curve Resolution to the microarray data allowed to obtain results of the same quality than the results obtained by standard microarray cluster analysis methods. An easier interpretation either of the sample clustering and of the co-expressed gens is possible. Use of the center of the final clusters of the k-means methods as initial estimation for MCR-ALS has been shown to improve quality of the final solutions.

Further interpretation of the obtained MCR gene profiles from a biomedical point of view is been carried out at present.

REFERENCES

[1]. **Microarray. Gene Expression Data Analysis.** Causton, H. C.; Quackenbush, J.; Brazma, A. (2003), (Blackwell, Oxford, UK)
[2]. Ross, D T et al.. **Systematic variation in gene expression patterns in human cancer cell lines.** Nature Genetics (2000), 24(3), 227-235.
[3]. Crescenzi, M; Giuliani, A. **The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data.** FEBS Letters (2001), 507(1), 114-118.
[4]. **Handbook of Chemometrics and Qualimetrics.** Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J.; Editors. (1997), (Elsevier, Oxford, UK)
[5]. Tauler, R. **Multivariate curve resolution applied to second order data.** Chemometrics and Intelligent Laboratory Systems (1995), 30(1), 133-46.

ACKNOWLEDGEMENTS

J.Jaumot acknowledges a PhD Grant from the Spanish Ministerio de Educación y Ciencia. This research was supported by the Spanish MEC (BQU2003-00191) and the Generalitat de Catalunya (2003SGR0056).