BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Tomul LVII (LXI), Fasc. 6, 2011
Secția
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ

UNSUPERVISED ANALYSIS OF MICROARRAY DATA

BY

RAUL MĂLUȚAN[1,*], PEDRO GOMEZ VILDA[2] and MONICA BORDA[1]

[1]Technical University of Cluj-Napoca
[2]Universidad Politécnica, Madrid, Spain

Received, May 31, 2011
Accepted for publication: July 19, 2011

Abstract. The proposed methods in this paper are based on a measurement of the expression level estimation procedures, considering also the statistical considerations from the hybridization process by the definition of an inner-product metric space. This helped in improving the estimation reliability and provided us a framework for de-noising microarray data *prior* to their use in unsupervised clustering analysis.

Key words: microarray; independent component analysis; unsupervised clustering.

1. Introduction

Current genomic microarray technology has become an advanced testing procedure for many different fields related with or supported by Functional Genomics. Microarray technology usage has experienced a certain explosion in the past years, well defined as a "gold rush" in a parallel metaphor (Knudsen, 2004).

This technology, conceived initially as a parallel implementation of well known techniques as Northern and Southern Blotting, has become an

* Corresponding author: *e-mail*: raul.malutan@com.utcluj.ro

available and easy-to-use standard procedure for the estimation of gene expression levels. Nevertheless important challenges have still to be faced to provide it with the desirable reliability levels required for its proper use. Many are the factors which result in unreliable estimates of expression levels, which are to be solved more by the microarray engineer or statistician than by the expert in Genomics. These are known as *microarray data processing challenges*.

The importance of side fields of knowledge as Signal and Image Processing, Pattern Recognition, Statistical Data Analysis, or Automata Theory in relation with microarray data processing challenges have not completely yielded their enormous potential in solving problems as microarray image enhancement, segmentation, correction, gridding, data analysis, reliable expression estimation in relation with hybridization dynamics, etc. Others have to see with data interpretation, dimensionality reduction, cluster analysis, function prediction, etc. Summarizing, the present paper will describe a combined method of microarray data processing: data correction by Independent Component Analysis followed by data clustering by unsupervised algorithms.

2. Independent Component Analysis of Microarray Data

The oligonucleotide microarrays are synthesized following a technique quite similar to VLSI microchip fabrication – photolithography. Firstly, the specific genes to be included in the microarray are selected and from these specific strands of 25 nucleotides are chosen. On a prepared glass slate a map of dots is drawn by photolithography using a mask for the first based to be deployed (*A*, *C*, *T* or *G*). The process is repeated for the four bases and for the 25-mer layers. The result is a test surface including an artificially built matrix of active *m*RNA strands. For each gene segment two *m*RNA strands are imprinted on the microchip: the one referred as Perfect Match (PM) is composed of a sequence of 25 bases specific to the segment reproduced; the second segment for the same gene is identical, except in its central base (the 13^{th} one), which is switched to its complementary base, this being referred as the MisMatch (MM) segment. Accordingly to hybridization laws it is expected that PM and MM would express specific and non specific hybridization, respectively. Expression estimating algorithms would have to take into account intensity levels for PM and MM to estimate a final expression levels for gene in question (Irizarry *et al*., 2003).

The detection of each gene depends on the multichannel differential expression of perfectly matched segments against mismatched ones. Considering this one can state that the amount of target species hybridized to a given sample, either PM or MM given as $h_{i,k}^{p}$ and $h_{i,k}^{m}$, respectively, at a specific spot in coordinates *x, y* of microarray could be given by

$$\begin{cases} h_{i,k}^{p} = \rho\left(s_{i,k}, x, y\right) p_t\left(s_{i,k} \left| z_{i,k}^{p} \right.\right), \\ h_{i,k}^{m} = \rho\left(s_{i,k}, x, y\right) p_t\left(s_{i,k} \left| z_{i,k}^{m} \right.\right), \end{cases} \tag{1}$$

as a function of the spatial distribution density of the target species, $\rho\left(s_{i,k}, x, y\right)$ and the conditioned probabilities of hybridization in time, $p_t\left(s_{i,k} \left| z_{i,k}^{m,p} \right.\right)$.

With these definitions a relationship between PM and MM expressions can be given by a proportionality parameter between both expression levels

$$\lambda_i = \frac{\left\| h_i^m \right\| \cos\beta_i}{\left\| h_i^p \right\|} = \frac{\left\langle h_i^m, h_i^p \right\rangle}{\left\| h_i^p \right\|^2} = \frac{\sum_{k=1}^{K} p_t(s_{i,k} \mid z_{i,k}^p) p_t(s_{i,k} \mid z_{i,k}^m)}{\sum_{k=1}^{K} p_t^2(s_{i,k} \mid z_{i,k}^p)}, \tag{2}$$

where β_i is the angle composed by the two expression vectors, $h_{i,k}^p$ and $h_{i,k}^m$.

The parameter defined in expression (2) helps in estimating the accuracy of the hybridization process. A derived parameter may serve to measure the orthogonality between PM and MM hybridization namely

$$\gamma_i = 1 - \cos^2 \beta_i. \tag{3}$$

The data in Table 1 give a good overview of how important are the corrections to be made during the pre-processing stages *prior* to pattern recognition and classification.

Table 1

Hybridization Reliability Estimation for Different Test Microarrays

Value of γ	HG-U133 chip	LatinSquare chip	MG_U74Av2 chip	MOR chip
< 0.05	11%	7%	24%	8%
0.05 ≤ γ < 0.1	17%	15%	27%	11.5%
0.1 ≤ γ < 0.5	68%	74%	47%	71.5%
0.5 ≤ γ	4%	4%	2%	9%
Total samples	22,283	22,300	12,488	1,824

A very relevant fact is that most of the PM–MM sample sets are not reliably expressed in the microarrays analysed, as their orthogonality factor (γ) is relatively high. Although this situation may be seem dramatic, current algorithmics cope the problem resourcing different methods, as removing pairs where the MM goes over the MM or similar.

One possibility of correction can be Independent Component Analysis

(ICA) (Măluțan *et al*., 2010). ICA allows us to better understand data in complex and noisy environments. It can separate the patterns in which we are interested from independent other effects like random sample variations or biological patterns unrelated to the subject of investigation. The technique has the potential of significantly increase the quality of the resulting data, and improve the biological validity of subsequent analysis (Lee & Batzoglou, 2003).

In our case it may be demonstrated (Măluțan *et al*., 2010) that the co-linear and orthogonal components, h_i^c and h_i^o, are already uncorrelated transformations of the original observations $h_{i,k}^p$ and $h_{i,k}^m$

$$h_i^c = \lambda_i h_{i,k}^p, \;\; h_i^o = h_{i,k}^m - h_i^c = h_{i,k}^m - \lambda_i h_{i,k}^p. \tag{4}$$

The working hypothesis is based on the assumption that the uncorrelated observations, h_i^c and h_i^o, are due to the linear combinations of unknown independent sources, s_i^a and s_i^b, namely

$$\begin{bmatrix} h_i^{cT} \\ h_i^{oT} \end{bmatrix} = A \begin{bmatrix} s_i^{aT} \\ s_i^{bT} \end{bmatrix}, \tag{5}$$

where the superscript *T* expresses the transpose (a row vector). Assuming the existence of an inverse, *W*, to *A*, the underlying sources or process correlates may be unleashed as

$$\begin{bmatrix} \hat{s}_i^{aT} \\ \hat{s}_i^{bT} \end{bmatrix} = W \begin{bmatrix} h_i^{cT} \\ h_i^{oT} \end{bmatrix}. \tag{6}$$

Once the estimates of the underlying sources, $\hat{s}_i^a$ and $\hat{s}_i^b$, as well as the combinations matrix $\hat{A}$ and the inverting matrix, *W*, are evaluated the corresponding re-estimates of the orthogonal and co-linear vector may be also evaluated

$$\begin{bmatrix} \hat{h}_i^{cT} \\ \hat{h}_i^{oT} \end{bmatrix} = \hat{A} \begin{bmatrix} \hat{s}_i^{aT} \\ \hat{s}_i^{bT} \end{bmatrix} \tag{7}$$

and from them, the re-estimated hybridization vectors it results

$$\hat{h}_i^p = \frac{\hat{h}_i^c}{\lambda_i}, \;\; \hat{h}_i^m = \hat{h}_i^o + \hat{h}_i^c. \tag{8}$$

The result of applying the above procedure to a set of unreliably expressed PM–MM test set is shown in Fig. 1.

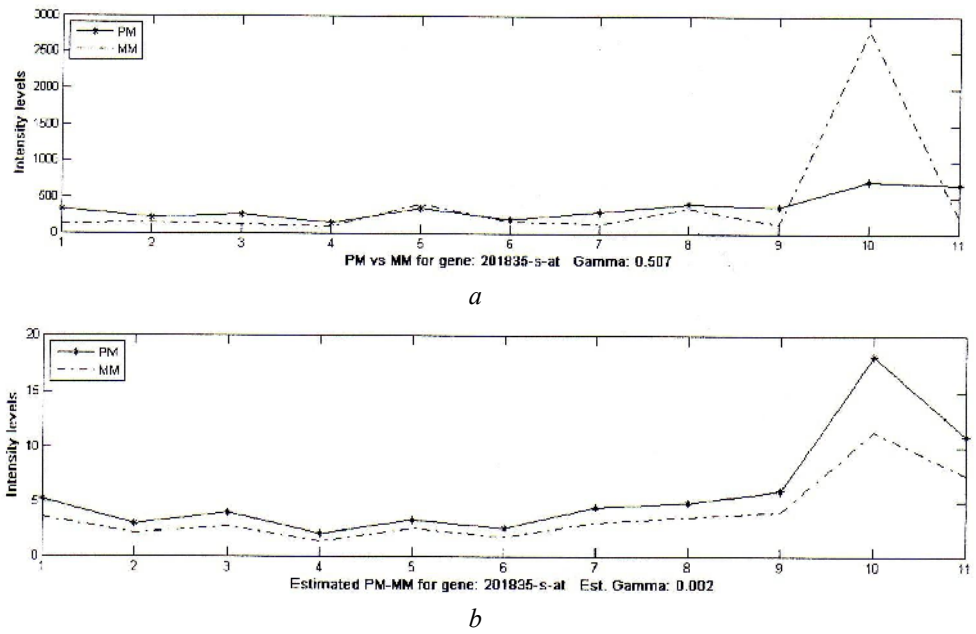

Fig. 1 – Application of ICA to microarray data to solve unknown sources and latent variables; *a* – unreliable expressed PM–MM sample sets; *b* – estimated PM–MM sample sets with gamma factor reduced from 0.507 to 0.002.

For a microarray database (CNIO) composed of three experiments, each with 22238 genes, from an average number of 14,603 unreliable labeled sample sets, only an average number of 4,115 sample sets showed an improving, this means a new computed $\gamma < 0.1$; and from an average number of 779 very unreliable labeled sample sets, the average number of corrected data is 474, as shown in Table 2.

Table 2

Number of ICA Corrected Unreliable Samples for a Microarray Database

Experiment name	Genes	Unreliable	Very unreliable	Corrected	
				Unreliable	Very unreliable
Tumores_A16T2(C)HG-U133A	22,283	15,613	1,088	4,320	677
Tumores_A23N(55)HG-U133A	22,283	14,980	599	4,249	349
Tumores_A19T(A)HG-U133A	22,283	15,076	842	4,112	495

3. Unsupervised Clustering of Microarray Data

The corrected data is next transferred to a gene expression matrix that will be analysed (Parmigiani *et al*., 2003) in order to extract some knowledge about the underlying biological processes. A gene expression matrix usually has the rows corresponding to genes from an experiment and the columns corresponding to different experiments. If one finds that two rows are similar, it can be assumed that the genes corresponding to the rows are co-regulated and functionally related, and by comparing two columns it can found which genes are differentially expressed in each experiment.

For the microarray data the most suitable clustering methods are unsupervized ones, because we cannot observe the (real) number of clusters in the data. In general, we can apply the cross-validation methods to a range of numbers of clusters in *k-means* or *Expectation–Maximization (EM)* clustering, and determine an estimate of optimal number of clusters from the data. Roman (2010) has clustered several databases, including microarray databases, and for each unsupervized clustering algorithm an optimal number of clusters were determined.

In our case, after the ICA correction the analysed microarray database was reduced from 22,283 to 11,490 genes. These genes were subject to clustering using the EM algorithm based on Gaussian mixture models (Bishop, 2006).

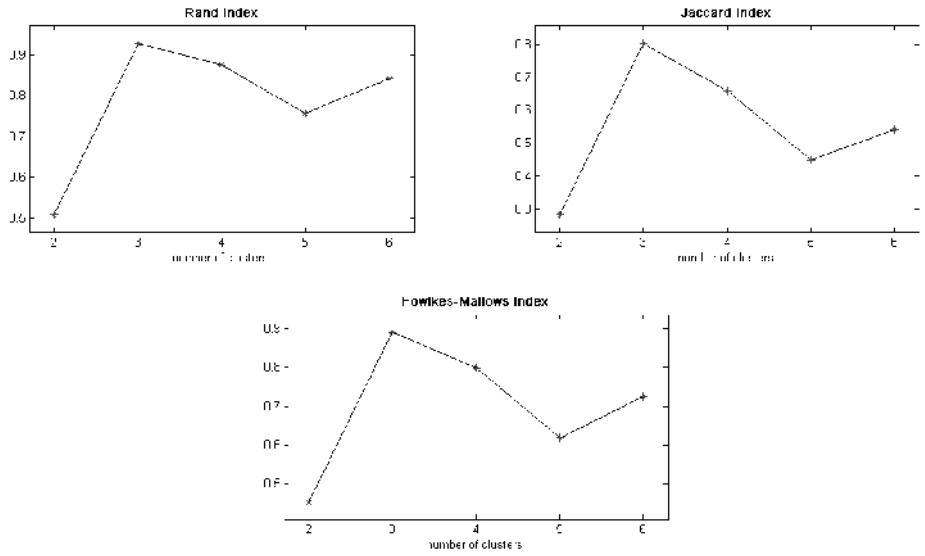

Fig. 2 – Cross-validation methods for determining the optimal number of clusters in the case when the EM algorithm was used.

For determining the optimal numbers of clusters for the current microarray data cross-validation methods were applied after a blind clustering

of the data. The methods included the usage of the external indexes Rand, Jaccard and Fowlkes-Mallows (Rendón *et al*., 2011; Costa *et al*., 2004), which assess the similarity between different partitions of the same dataset. The indexes produce a result in the range [0, 1] for each above mentioned index, where a value of 1 for a certain number of clusters means that is the optimal number of clusters for that data. From a maximum number of 6 clusters, the optimal number of clusters for all indexes was found to be 3, as it can be seen in Fig. 2.

Once the number of clusters was set we rerun the EM algorithm with 3 clusters. The results of clustering are shown in Fig. 3, with a probability distribution of the numbers of genes in each cluster of 0.5437% for the first cluster, 0.3774% for the second cluster and 0.0878% for the third cluster.

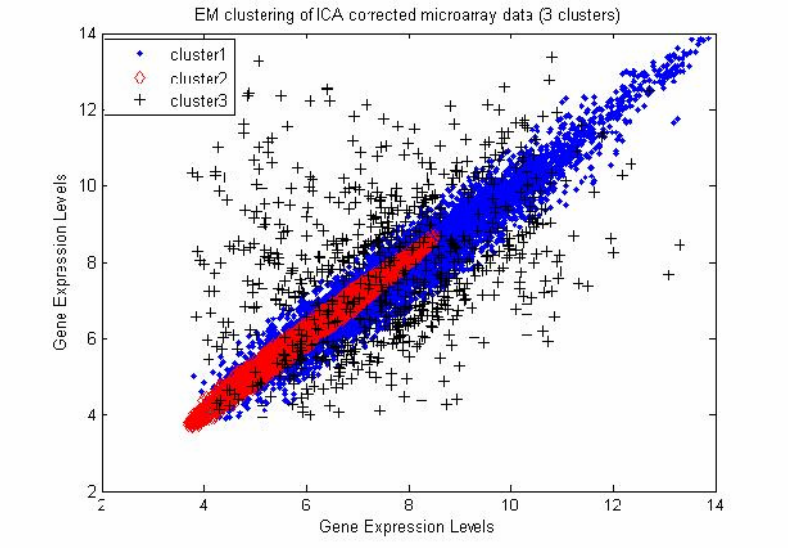

Fig. 3 – The results of EM clustering when 3 clusters were used; the first cluster has the largest number of genes, while the third one has only few genes.

4. Conclusions

The unsupervized analysis of the microarray data was done using different methods with different purposes. Firstly we used Independent Component Analysis as a technique powerful enough to specifically correct deviations produced by unknown factors by extracting them and using their trace to be removed from the observations. Then we used a unsupervized clustering, but not only for clustering the remaining gene expression levels, but also for determining an optimal number of clusters for a microarray data for which no *prior* information was given.

Acknowledgment. This paper was supported by the project "Development and Support of Multidisciplinary Postdoctoral Programmes in Major Technical Areas of National Strategy of Research – Development – Innovation" 4D-POSTDOC, contract no. POSDRU/89/1.5/S/52603, project co-funded by the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013.

REFERENCES

Bishop C.M., *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006.

Costa I. G.,de Carvalho F., de Souto M., *Comparative Analysis of Clustering Methods for Gene Expression Time Course Data*. Genet. a. Molec. Biol., **27**, *4*, 623-631 (2004).

Irizarry R.A. *et al.*, *Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data*. Biostatistics, **4**, *2*, 249-264 (2003).

Knudsen S., *Guide to Analysis of DNA Microarray Data*. J. Wiley and Sons Inc., NY, 2004.

Lee S.I., Batzoglou S., *Application of Independent Component Analysis to Microarrays*. Genome Biol., **4**, R76.1 - R76.21 (2003).

Măluțan R., Gómez P., Borda M., *Independent Component Analysis Algorithms for Microarray Data Analysis*. Intell. Data Anal. J., **14**, *2*, 193-206 (2010).

Parmigiani G., Garrett E.S., Irizarry R.A., Zeger S.L., *The Analysis of Gene Expression Data*. Springer-Verlag, New York, 2003.

Rendón E., Abundez I., Arizmendi A., Quiroz E. M. *Internal versus External Cluster Validation Indexes*. Internat. J. of Comp. a. Commun., **5**, *1*, 27-34 (2011).

Roman A., *Unsupervized Methods for Data Clustering*. B.Sc. Diss., Techn. Univ. of Cluj-Napoca, 2010.

* * * Centro Nacional de Investigaciones Oncologicas (CNIO). http://www.cnio.es/ing/

ANALIZA NESUPERVIZATĂ A DATELOR MICROARRAY

(Rezumat)

Metodele propuse în această lucrare se bazează pe estimări ale măsurătorilor efectuate pentru nivelul de exprimare genetică pentru datele microarray. S-au luat în considerare şi efectele introduse de procesul de hibridizare. Aceste măsurători au creat un cadru pentru estimări fiabile, dar şi pentru preprocesarea datelor microarray, etapă utilizată înaintea clasificărilor de tip nesupervizat.