

Microarray Data Statistical Analysis on Myc Gene Network

Analisi Statistica di Dati di Microarray: la Rete Genica di Myc

Marco Scarnò

CASPUR, Roma – E-mail: mscarno@caspur.it

Sergio Nasi

IBPM CNR, Università La Sapienza, Roma – E-mail: Sergio.Nasi@uniroma1.it

Donatella Sforzini

CASPUR, Roma – E-mail: d.sforzini@caspur.it

Riassunto: I chip di DNA consentono di monitorare l'espressione di migliaia di geni simultaneamente e costituiscono una importante opportunità per le scienze bio-mediche. Per loro complessità, tuttavia, i dati derivati da tali supporti pongono una serie di problemi dal punto di vista dell'analisi statistica: ciò che presenta le maggiori difficoltà è l'attribuzione del significato biologico e delle condizioni funzionali connesse alla classificazione dei geni. Una stima diretta della relazione tra variabili (condizioni sperimentali) e osservazioni (geni) può essere ottenuta con l'Analisi delle Corrispondenze, tramite la loro proiezione in uno spazio compromesso. Analizzando i profili di espressione genica di cellule (fibroblasti) in presenza di un mutante Myc dominante negativo, abbiamo verificato che l'analisi delle Corrispondenze è un valido ausilio per la classificazione di dati di Microarray.

Keywords: Microarrays Data Analysis, Correspondence Analysis.

1. Introduction

Gene expression profiling by DNA microarrays is increasingly used for functional studies of living cells and for clinical diagnostics. Microarray data are represented as a statistical table, in which relationships among rows (genes) and columns (hybridization experiments) must be found and evaluated. Data analysis methods are aimed at defining groups of genes with similar expression profiles in a given set of conditions. A variety of statistical algorithms have been used, ranging from clustering algorithms to supervised learning approaches such as discriminant analysis and neural networks. Hierarchical Clustering algorithms group similar objects based on the evaluation of a distance metric, usually the Euclidean distance, between gene expression profiles. Although suited for analysing associations between genes in time course studies, they proved to be relatively unsuccessful for inferring gene function or for comparison between experiments. An alternative is provided by projection algorithms, such as Principal Component Analysis (PCA) and Correspondence Analysis (CA). The underlying idea of CA is to study the level of dependence between rows and columns (genes and hybridisations in microarray experiments) by first embedding both in an artificial space with a small number of dimensions, and then scoring similarity by proximity in the projection space. In statistical terms, rows and columns are

independent when row values do not change in correspondence of different column values. Therefore, a *perfect independence* condition holds when all genes display the same expression value in all hybridisations. Vice versa, a *perfect dependence* condition is present when each gene has an expression value different from zero in only one hybridisation. Obviously these are extreme cases, but Correspondence Analysis can measure the *distance* between the *perfect independence* condition and the real case and evaluate the influence of each column to the determination of each row, or vice versa. In CA the original data can be reconstructed at a desired level of *information loss*, making it possible to *drop* the part of information supposedly due to experimental errors, although this may not always be easy to evaluate in biological experiments. In this work, we report the analysis of microarray experiments on gene regulation by Myc, a protein that binds to DNA and controls the expression of a large set of genes. As Myc activity is commonly altered in human cancer, identifying and classifying these genes is thought to be important for understanding the pathways that control the behaviour of normal and tumour cells.

2. Results, methods and discussion

The expression profile of about 8799 genes was measured on DNA oligonucleotide arrays (RGU34A Chip, from Affymetrix) in Rat1 fibroblast cells stimulated to proliferate by serum addition to the culture medium. Three time points were evaluated: 0, 90 and 240 minutes. For each time point, data were collected both in presence and absence of Omomyc, a mutant protein that interferes with Myc biological activity (Soucek et al. (2002)). Raw fluorescence intensity data obtained by laser scanning of the hybridisation experiments were normalized and processed with the GeneChip Microarray Analysis Suite version 4.0 (MAS 4) software. Comparison of different samples was performed in two steps. The first one is an absolute analysis that utilizes intensity data from the entire probe array for background and noise calculations, for assessing the absolute expression level of any gene represented by the array probe sets, and for computing a variety of parameters used by a decision matrix to determine whether a transcript is Present (P), Marginal (M) or Absent (A). In the second step a single condition or sample is taken as reference, with the aim of identifying differences in gene expression profile between experimental and reference arrays for every probe set. The most relevant variable for comparative analysis is the Fold Change (FC), an indicator of the relative change of expression of each gene represented on the probe array. Fold Change depends on several factors such as average level of gene expression, array type and noise of the experiment. In the MAS 4 software, it has the form:

$$FC = \frac{\overline{Diff_{exp}} - \overline{Diff_{base}}}{\max\left[\min\left(\overline{Diff_{exp}}, \overline{Diff_{base}}\right), Q_M * Q_C\right]} + \begin{cases} +1 & \text{if } \overline{Diff_{exp}} \geq \overline{Diff_{base}} \\ -1 & \text{if } \overline{Diff_{exp}} < \overline{Diff_{base}} \end{cases}$$

where:

$$\overline{Diff_{exp}} = \frac{\sum (PM - MM)}{Pairs \text{ in Avg}}$$

is the Average Difference calculated on the experiment and $\overline{Diff_{base}}$ is the same indicator calculated on the reference. FC is a positive number if gene expression (transcript abundance) is increased, and a negative number if it is decreased with respect to the reference sample. In microarray experiments, true differences in gene expression can be hidden by the potential presence of a large number of random and systematic errors. Other difficulties are the consequence of the small number of experiments relative to the large number of variables. Only differences in expression exceeding an empirical threshold, defined by the presence of noise, are considered statistically significant. Data above the threshold, usually set at a FC value ranging from two to three, are kept for further analysis (data filtering). The threshold in our case was set to +3 for up-regulated genes and -3 for the down-regulated ones, the condition of cells without Omomyc at time 0 serving as reference.

Our interest was directed at identifying genes that were either significantly decreased or increased in the presence of Omomyc, as compared to cells without Omomyc at the same time points, and at assessing how well CA would perform in this task. We focused the statistical analysis on four conditions, defined by the presence (+) or absence (-) of the Omomyc mutant at the two time points 0 and 90 min, and proceeded in the following way. Genes were manually sorted according to the robust analysis rules indicated by the array manufacturer (Affymetrix Data Analysis Fundamentals), using the parameters calculated by the MAS 4 software. At the same time, Correspondence Analysis was performed on the same data matrix, via routine present in the SAS package. As the GeneChips employed contained 8799 probe sets, the original data were in the form of a 8799×3 matrix, in which rows represent genes (more precisely probe sets) and columns experiments: the value of each cell is the Fold Change of a single probe set with respect to the reference condition. Since CA algorithms require a data matrix of positive numbers, and Fold Change can take negative values, we had to split the columns into two parts made only of positive values and representing, respectively, activated and repressed genes. Therefore, CA was performed on a 8799×6 data matrix. Values were projected in a three dimensional space, which allowed an easy visualization while still maintaining around 80% of the global Inertia of the data set. Each point (row or column value) in the projection space was characterized by the vector connecting it to the origin, and the association between points was evaluated through the cosine of the angle formed by their respective vectors into the projection space. In particular a gene will be associated to the conditions corresponding to the largest value of the cosine itself. We found that the experimental conditions are quite well separated by the CA (Fig. 1). The results obtained through CA were then compared with those of the manual selection through robust analysis. We observed that 89% of over 200 genes identified by the manual procedure as decreased at 90 min in the presence of Omomyc were also detected by the CA. The same comparison on genes increased by the induction of Omomyc at 90 min (around 60 genes) gave again a remarkable fit between the two methods, with about 87% genes in common.

In conclusion, we find that Correspondence Analysis constitutes a valuable tool for revealing associations between genes and conditions. Moreover this technique appears to be more straightforward than other statistical methods, as also reported by other authors (Fallenberg et al., 2001, Wouters et al., 2003). Easily evaluating such associations may be of great value in the analysis of clinical samples, where it is important to find out which genes better characterize a given pathological condition.

Scatterplot Correspondence Analysis

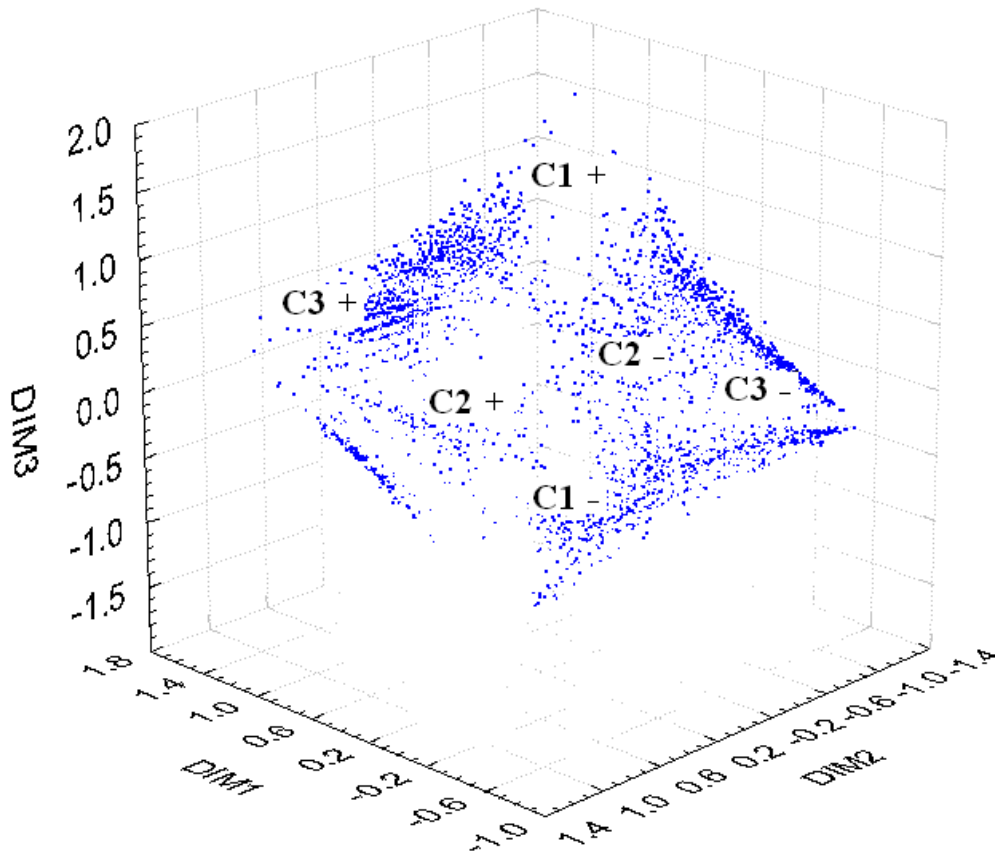


Figure 1: Scatterplot of Correspondence Analysis. Points *c1-c3* correspond to the tree time condition 0, 90 and 240 min. Signs + and – characterize the positive or negative values of the Fold Change, i.e. genes whose expression is increased (+) or decreased.

References

- Eisen M. B., Spellman P. T., Brown P. O. & Botstein, D (1988) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci.*,95, 14863-14868, USA.
- Fellenberg K., Hauser N. C., Benedikt B., Neutzer A., Hoheisel J. D. & Vingron M. (2001) Correspondence Analysis applied to microarray data, *Proc. Natl. Acad. Sci.*,98, 10781-10786, USA.
- Soucek L., Jucker R., Panacchia L., Tatò F. & Nasi S. (2002) Omomyc, a potential Myc dominant negative, enhances Myc-induced apoptosis, *Cancer Research*,62, 3507-3510.
- Wouters L., Hinrich Göhlmann W., bijnens L., Stefan Kass U., Molenberghs G., Lewi P. J. (2003) Graphical Exploration of Gene Expression Data: A Comparative Study of Three Multivariate Methods, *Biometrics*,59, 1131-1139.