BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 9 2006, pages 1096–1102
doi:10.1093/bioinformatics/btl056

Gene expression

maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments

Ana Conesa[1,†,*], María José Nueda[2,†], Alberto Ferrer[3] and Manuel Talón[1]

[1]Centro de Genómica. Instituto Valenciano de Investigaciones Agrarias, Apartado Oficial 46113, Moncada, Valencia, Spain, [2]Departamento de Estadística e Investigación Operativa. Universidad de Alicante. Apartado 03080, Alicante Spain and [3]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia, Apartado 46022, Valencia, Spain

Received on November 9, 2005; revised on February 1, 2006; accepted on February 10, 2006
Advance Access publication February 15, 2006
Associate Editor: David Rocke

ABSTRACT

Motivation: Multi-series time-course microarray experiments are useful approaches for exploring biological processes. In this type of experiments, the researcher is frequently interested in studying gene expression changes along time and in evaluating trend differences between the various experimental groups. The large amount of data, multiplicity of experimental conditions and the dynamic nature of the experiments poses great challenges to data analysis.

Results: In this work, we propose a statistical procedure to identify genes that show different gene expression profiles across analytical groups in time-course experiments. The method is a two-regression step approach where the experimental groups are identified by dummy variables. The procedure first adjusts a global regression model with all the defined variables to identify differentially expressed genes, and in second a variable selection strategy is applied to study differences between groups and to find statistically significant different profiles. The methodology is illustrated on both a real and a simulated microarray dataset.

Availability: The method has been implemented in the statistical language R and is freely available from the Bioconductor contributed packages repository and from http://www.ivia.es/centrogenomica/bioinformatics.htm

Contact: aconesa@ivia.es; mj.nueda@ua.es

1 INTRODUCTION

A general approach in experimental life science research is to monitor the evolution over a period of time of biological phenomena as a response to specific stimuli. From a functional genomics point of view, the genome-wide study of temporal variations in gene expression aims to understand the molecular basis that control biological processes. Microarray technology allows to monitor the expression levels of thousands of genes simultaneously [see Draghici (2003) for an overview] and is therefore a very useful methodology to address the analysis of gene expression changes over time (microarray time course, MTC). The design of a typical time-course experiment often includes a number of experimental treatments that are monitored through a relatively small (<6) number of time points. The researcher is then interested in detecting biologically meaningful gene expression trends and in spotting differences between the various experimental groups.

Clustering methods, habitually used for the study of gene expression profiles, have been applied to the analysis of time-course data (Spellman *et al*., 1998; Lukashin *et al*., 2001). These methods cluster gene expression profiles on the basis of a distance metric and are valuable tools for the visualization of these data and for identifying groups of co-regulated genes (Draghici, 2003; Speed, 2003). In some cases, a statistical assesment for cluster significance is provided along with the clustering approach (Kerr and Churchill, 2001; Herrero *et al*., 2001), but in general these techniques do not offer an adequate framework to asses statistically significant trend differences between conditions. Furthermore, when a large number of genes is present in the dataset the interpretation of clustering results can be problematic. Therefore, it seems more convenient to apply first a statistical procedure to identify those genes with significant expression changes and subsequently divide the gene selection into clusters to visualize the results.

Traditional statistical methods (*t*-statistic tests, ANOVA, etc.) have been applied to microarray data to identify differentially expressed genes (Pan, 2002; Kerr *et al*., 2000; Wolfinger *et al*., 2001). Refinements of these methods that take into account particular properties of gene expression data are now available. Some popular examples are SAM (Significance Analysis of Microarrays, Tusher *et al*., 2001) and LIMMA (Linear Models for Microarray Data, Smyth, 2004). These methods, although powerful and easy to use, are focused mainly on pairwise comparisons and their application to microarray time courses, specially when multiple series are present, might be tedious and uneffective to capture the dynamic nature of this type of data.

The statistical analysis of microarray time-course data has been reviewed by Bar-Joseph (2004). A large number of currently available methods is devoted to the identification and clustering of gene expression patterns, and for the deciphering of gene regulatory networks (references in Bar-Joseph, 2004; Peddada *et al*., 2003; Luan and Li, 2003; Liu *et al*. 2005; Ernst *et al*. 2005 and Beal *et al*. 2005). However, few methodologies can be found that address the problem of finding statistical profile differences between

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

experimental groups. Bar-Joseph *et al*. (2003) obtained a selection of differentially expressed genes between two cell-cycle microarray datasets by computing a difference measure between the continuous representations of the two time series expression data. This method can be successfully applied to the analysis of long time series (>10 time points) but its adequateness for shorter time-course experiment is not clear (Bar-Joseph, 2004). ANOVA-based models have also been proposed (Park *et al*., 2003). ANOVA can easily model multilevel factors and their interactions. However, when analysing models containing quantitative variables or experiments with unbalanced designs traditional ANOVA procedures are not appropriate and specific modifications have to be incorporated. Regression approaches appear to be a more straightforward and flexible solution for the analysis of this type of data. Regression methods treat time as a quantitative variable, and therefore not only differentially expressed genes can be detected, but also changes in trends can be discovered and their magnitude can be studied by analysing the coefficients of the model. A regression model approach was used by Xu *et al*. (2002) to identify differential gene profiles in an inducible transgenic model. Their method introduced specific variables in the regression to capture particular properties of the data under study. This tailor-made approach can be very useful to evaluate specific gene expression behaviours but it implies redefining the variables for other biological systems. In this work we propose a general regression-based approach for the analysis of single or multiple microarray time series. This methodology, named maSigPro (microarray Significant Profiles) is a two-step regression strategy where model parameters have to be adjusted according to the data under study and the specific interests of the researcher.

The proposed method has been successfully applied to several experiments. In this work the procedure is illustrated both on simulated data and a public domain toxicogenomics dataset. The methodology has been made available as an R package.

2 METHODS

2.1 Definition of the model

In the problem we are considering there are normally two or more variables of interest. One of them is typically the time, which is a quantitative variable (in the type of experiments considered for this approach, time is usually the independent variable, however the methodology would accept as well other experimental continuous variables, such as a quantified physiological parameter). The other variables are usually qualitative variables (e.g. different treatments, strains, tissues, etc.) and represent the experimental groups for which temporal gene expression differences are sought. For clarity in the exposition, only one qualitative variable or factor will be considered here.

Let there be I experimental groups described by the qualitative variable evaluated at J time points for each particular condition ij ($i = 1, \ldots, I$ and $j = 1, \ldots, J$). Assume that gene expression is measured for N genes in R_{ij} replicated hybridizations.

We define $I-1$ dummy variables (binary variables) to distinguish between each group and a reference group (Table 1).

Let y_{ijr} denote the normalized and transformed expression value from each gene in the situation ijr ($r = 1, \ldots, R_{ij}$). To explain the evolution of y along the time (T) we consider the following polynomial model, where simple time effects and interactions between the dummies and the time have been modelled. In principle, the maSigPro methodology allows a polynomial model of $J-1$ degree as the model described in Equation (1).

Table 1. Definition of experimental groups with dummy variables

Group	D_1	D_2	$\cdots$	D_{I-1}
1 (Ref. group)	0	0	0	0
2	1	0	0	0
3	0	1	0	0
...	0	0	...	0
I	0	0	0	1

$$\begin{aligned} y_{ijr} = {} & \beta_0 + \beta_1 D_{1ijr} + \cdots + \beta_{(I-1)} D_{(I-1)ijr} \\ & + \delta_0 T_{ijr} + \delta_1 T_{ijr} D_{1ijr} + \cdots + \delta_{(I-1)} T_{ijr} D_{(I-1)ijr} \\ & + \gamma_0 T_{ijr}^2 + \gamma_1 T_{ijr}^2 D_{1ijr} + \cdots + \gamma_{(I-1)} T_{ijr}^2 D_{(I-1)ijr} \\ & \cdots \\ & + \lambda_0 T_{ijr}^{J-1} + \lambda_1 T_{ijr}^{J-1} D_{1ijr} + \cdots + \lambda_{(I-1)} T_{ijr}^{J-1} D_{(I-1)ijr} + \varepsilon_{ijr} \end{aligned} \quad (1)$$

$\beta_0, \delta_0, \gamma_0, \ldots, \lambda_0$ are the regression coefficients corresponding to the reference group. $\beta_i, \delta_i, \gamma_i \ldots \lambda_i$ are the regression coefficients that account for specific differences (linear, quadratic, cubic, etc.) between the $(i+1)$-th group profile and the first group (reference) profile, $i = 1, \ldots, I-1$. ε_{ijr} is the random variation associated with each gene in each hybridization ijr owing to all sources other than those that have already been incorporated into the model.

This model defines implicitly as many models as experimental groups. For example, the model for the first group is $y_{1jr} = \beta_0 + \delta_0 T_{1jr} + \gamma_0 T_{1jr}^2 + \cdots + \lambda_0 T_{1jr}^{J-1} + \varepsilon_{1jr}$, since in this group all the dummies are 0; and for the second group is $y_{2jr} = (\beta_0 + \beta_1) + (\delta_0 + \delta_1) T_{2jr} + (\gamma_0 + \gamma_1) T_{2jr}^2 + \cdots + (\lambda_0 + \lambda_1) T_{2jr}^{J-1} + \varepsilon_{2jr}$. In this example $\beta_1, \delta_1, \gamma_1, \ldots, \lambda_1$ measure the differences between the second and first (reference) groups related to linear, quadratic, etc. and $(J-1)$-th time order effects; respectively.

2.2 First regression model: gene selection

The first step of the maSigPro approach applies the least-squares technique to estimate the parameters of the described general regression model for each gene. This means that we are testing the following null and alternative hypotheses:

$$\begin{aligned} H_0 : {} & \beta_1 = \cdots = \beta_{I-1} = \delta_0 = \delta_1 = \cdots = \delta_{I-1} = \gamma_0 = \gamma_1 = \cdots = \\ & \gamma_{I-1} = \cdots = \lambda_0 = \lambda_1 = \cdots = \lambda_{I-1} = 0 \\ H_1 : {} & \exists i / \beta_i \neq 0, (i = 1, \ldots, I-1) \\ & \vee \delta_i \neq 0 \vee \gamma_i \neq 0 \vee \cdots \vee \lambda_i \neq 0, (i = 0, \ldots, I-1) \end{aligned} \quad (2)$$

This first analysis generates N ANOVA tables as shown in Table 2, one for each gene. A gene with different profiles between the reference group and any other experimental group will show some statistically significant coefficient, and its corresponding regression model will be statistically significant. The P-value associated to the F-Statistic in the general regression model is used to select significant genes. This P-value is corrected for multiple comparisons by applying the linear step-up (BH) false discovery rate (FDR) procedure (Reiner *et al*., 2002). Therefore, genes with a FDR lower than a predetermined threshold will be selected.

2.3 Second regression step: variable selection

Once statistically significant gene models have been found, the regression coefficients of the models can be used to identify the conditions for which genes shows statistically significant profile changes. To do this, a new model is obtained only for selected genes, applying a variable selection strategy (stepwise regression, Draper and Smith, 1998). Stepwise regression is an iterative regression approach that selects from a pool of potential variables the 'best' ones (according to a specified criterion) to fit the available data. In this process, the statistical significance of the regression coefficients of the variables present in the model at each iteration is computed and only those

Table 2. ANOVA table. $\hat{y}$ is the predicted expression value, $\bar{y}$ is the average expression value and p is the number of variables in the model, (polynomial order +1) $I-1=JI-1$

Source	Sum of squares (SC)	Degrees of freedom	Mean square error	F-Statistic
Regression	$\mathrm{SCR}=\sum_{ijr}(\hat{y}_{ijr}-\bar{y})^2$	p	SCR/p	$\frac{(\mathrm{SCR}/p)}{\mathrm{SCE}/[\sum_{i,j}R_{i,j}-(p+1)]}$
Error	$\mathrm{SCE}=\sum_{ijr}(y_{ijr}-\hat{y}_{ijr})^2$	$\sum_{i,j}R_{ij}-(p+1)$	$\frac{\mathrm{SCE}}{\sum_{i,j}R_{ij}-(p+1)}$	
Total	$SCT=\sum_{ijr}(y_{ijr}-\bar{y})^2$	$\sum_{i,j}R_{ij}-1$		

Table 3. Results matrix of regression coefficients for the variable selection fit. Genes are given in rows and model parameters in columns. Regression coefficients associated to the same dummy variables are labelled with the same number. NA value for regression coefficients indicates that the variable was not statistically significant for that gene (under a given threshold, type I risk)

	1	2	3	⋯ I	1	2	3	⋯ I		1	2	3	⋯ I
GeneID	β_0	β_1	β_2	⋯ β_{I-1}	δ_0	δ_1	δ_2	⋯ δ_{I-1}	⋯	λ_0	λ_1	λ_2	⋯ λ_{I-1}
	Intercept	D_1	D_2	⋯ D_{I-1}	Time	Time × D_1	Time × D_2	⋯ Time × D_{I-1}	⋯	Time^{J-1}	$\mathrm{Time}^{J-1}\times D_1$	$\mathrm{Time}^{J-1}\times D_2$	⋯ $\mathrm{Time}^{J-1}\times \mathrm{D}_{I-1}$
Gene 1	β_{01}	β_{11}	NA	⋯ NA	NA	δ_{11}	NA	⋯ NA	⋯	NA	λ_{11}	NA	⋯ $\lambda_{(I-1)1}$
Gene 2	β_{02}	NA	NA	⋯ NA	δ_{02}	NA	δ_{22}	⋯ $\delta_{(I-1)2}$	⋯	λ_{02}	NA	λ_{22}	⋯ NA
Gene 3	NA	NA	NA	⋯ $\beta_{(I-1)3}$	NA	δ_{13}	NA	⋯ $\delta_{(I-1)3}$	⋯	NA	λ_{13}	NA	⋯ $\lambda_{(I-1)3}$
Gene 4	NA	β_{14}	β_{24}	⋯ $\beta_{(I-1)4}$	δ_{04}	NA	δ_{24}	⋯ $\delta_{(I-1)4}$	⋯	λ_{04}	NA	λ_{24}	⋯ NA
⋯	⋯	⋯	⋯	⋯ ⋯	⋯	⋯	⋯	⋯ ⋯	⋯	⋯	⋯	⋯	⋯ ⋯
Gene N	β_{0N}	NA	β_{2N}	⋯ $\beta_{(I-1)N}$	δ_{0N}	NA	δ_{2N}	⋯ NA	⋯	λ_{0N}	NA	λ_{2N}	⋯ $\lambda_{(I-1)N}$

variables with a P-value under a given threshold (type I risk) are maintained. In this case, applying FDR for multiple comparisons is not easy due to the fact that P-values associated to each coefficient vary as the model evolves. Therefore, we apply a threshold that must be fixed by the researcher. We recommend to correct the desired level of significance for the total possible number of variables in the model. The variables included in these new models will be those that indicate the differences in profiles. The maSigPro package provides different types of stepwise regression: backward, forward, stepwise backward and stepwise forward. This variable selection approach has a double effect: on one hand it provides the significant differences between experimental groups, and on the other hand, it generates an adequate regression model for the data. This implies that for each gene and experimental group, polynomial regressions of different degree (up to the maximum initially given in the formulation of the model) can be obtained. The method will therefore generate a matrix with so many rows as significant genes and so many columns as parameters in the complete regression model [Equation (1)]. This results matrix contains information (estimated coefficient and its P-value) for those variables that remained in the model of each gene. Table 3 is an illustrating example of such a results matrix.

This matrix provides the framework for selecting significant genes for each variable of the complete model and for each experimental group. For example, to find genes that have significant differences in group 2 respect to the reference group, those genes having statistically significant coefficients for the variables associated to the Dummy1 (D_1, Time$\times D_1,\ldots,\mathrm{Time}^{J-1}\times D_1$) must be selected, i.e. genes which have a significant β_1, $\delta_1,\ldots,\lambda_1$ coefficients (columns labelled as 2 in Table 3). In addition, the study of individual model variables allows focussing on the evaluation of specific pattern differences. For example, the analysis of the regression coefficients of the variable Time × D_1 allows the classification of genes for their different behaviour in the linear model component (i.e. induction or repression) of group 2 with respect to the reference group. The maSigPro package includes functions to easily perform different types of gene selection at this stage.

Until now, the goodness of fit (R-squared) of the new models has not been considered. This means that all significant genes are selected genes. The researcher might however be interested only in genes with clear trends as this may reflect biologically meaningful behaviours. In such case, maSigPro allows an additional gene selection step based on the R-squared value of the second regression model.

2.4 Visualization

The maSigPro package provides a number of functions for the visual analysis of the results. Individual plots of expression profiles by experimental group can easily be generated for each significant gene. Computed regression curves can also be superimposed to visualize the modelling obtained for the data. When the number of selected genes is large, cluster algorithms may be used to split the data into groups of similar expression patterns. maSigPro incorporates a number of traditional clustering algorithms to do so. These algorithms typically use gene expression data to compute clusters. In additional, maSigPro provides a clustering alternative that uses the estimated regression coefficients rather than the original data. This option will group genes on the basis of their statistically significant profile changes, discarding the noise of the data that has been removed by the estimated model. Once clusters have been obtained, maSigPro displays both the continuous expression profile along all experimental conditions and the average expression profile by experimental group for each cluster. The first representation helps to analyze the homogeneity of the clusters while the second provides a useful visualization of the between-groups differences for the genes of each cluster.

3 RESULTS

3.1 Case 1: toxicogenomics dataset

The maSigPro method has been applied to the analysis of a published dataset from a toxicogenomics study where the effect of the

Table 4. Treatments assigned to each slide. The indicated dye is the assigned to the sample

Slide	Treatment	Slide	Treatment	Group	Time
1	Cy3-UT-T6	28	Cy5-UT-T6	1	6
2	Cy3-UT-T24	29	Cy5-UT-T24	1	24
3	Cy3-UT-T48	30	Cy5-UT-T48	1	48
4	Cy3-CO-T6	31	Cy5-CO-T6	2	6
5	Cy3-CO-T24	32	Cy5-CO-T24	2	24
6	Cy3-CO-T48	33	Cy5-CO-T48	2	48
7	Cy3-LBB-T6	34	Cy5-LBB-T6	3	6
8,9,10	Cy3-LBB-T24	35–37	Cy5-LBB-T24	3	24
11	Cy3-LBB-T48	38	Cy5-LBB-T48	3	48
12	Cy3-MBB-T6	39	Cy5-MBB-T6	4	6
13–15	Cy3-MBB-T24	40–42	Cy5-MBB-T24	4	24
16–18	Cy3-MBB-T48	43–45	Cy5-MBB-T48	4	48
19–21	Cy3-HBB-T6	46–48	Cy5-HBB-T6	5	6
22–24	Cy3-HBB-T24	49–51	Cy5-HBB-T24	5	24
25–27	Cy3-HBB-T48	52–54	Cy5-HBB-T48	5	48

hepatotoxicant bromobenzene in rats was studied (Heijne *et al*., 2003). Rats were treated with three doses (low, medium and high) of bromobenzene dissolved in corn oil. In addition, there were two groups of rats without toxic treatment: an untreated rats group and a group treated only with the drug administration vehicle, corn oil. In total there were five groups denoted by the labels: UT (untreated), CO (corn oil), LO (low dose), ME (medium dose) and HI (high dose). Each individual RNA rat sample was co-hybridized with an external reference and the hybridizations were duplicated swapping the two labelling dyes. At different measurement time-points (6, 24 and 48 h) one to three rats were randomly selected from each treatment group. Individual rat RNA samples were co-hybridized against an external reference and hybridizations were duplicated swapping the two labelling dyes. This makes a total of 54 slides (Table 4) and 2665 genes available for statistical analysis. Data pre-processing included background subtraction, calculation of log2 ratios and Lowess normalization. In addition, a possible dye effect was estimated and removed from each gene.

In this example there are five experimental groups ($i = 1, \ldots, 5$), three time points ($j = 1, 2, 3$), two or six observations, $r = 1, \ldots, R_{ij}$ (2 or 6) for each case ij, and 2665 genes ($n = 1, \ldots, 2665$). The CO group was taken as reference group as this provides the true control for the treatments. Consequently, we defined four dummy variables D_{UT}, D_{LO}, D_{ME} and D_{HI} to introduce in the model the experimental groups in an analogous way as described in Table 1. We considered for each gene the model given in Equation (3) where linear and quadratic time effects and their interactions with the dummies have been modelled.

$$\begin{aligned} y_{ijr} = {} & \beta_0 + \beta_1 D_{(\mathrm{UT})ijr} + \beta_2 D_{(\mathrm{LO})ijr} + \beta_3 D_{(\mathrm{ME})ijr} + \beta_4 D_{(\mathrm{HI})ijr} \\ & + \beta_5 T_{ijr} + \beta_6 D_{(\mathrm{UT})ijr} \times T_{ijr} + \beta_7 D_{(\mathrm{LO})ijr} \times T_{ijr} \\ & + \beta_8 D_{(\mathrm{ME})ijr} \times T_{ijr} + \beta_9 D_{(\mathrm{HI})ijr} \times T_{ijr} + \beta_{10} T^2_{ijr} \\ & + \beta_{11} D_{(\mathrm{UT})ijr} \times T^2_{ijr} + \beta_{12} D_{(\mathrm{LO})ijr} \times T^2_{ijr} + \beta_{13} D_{(\mathrm{ME})ijr} \\ & \times T^2_{ijr} + \beta_{14} D_{(\mathrm{HI})ijr} \times T^2_{ijr} + \varepsilon_{ijr}. \end{aligned} \tag{3}$$

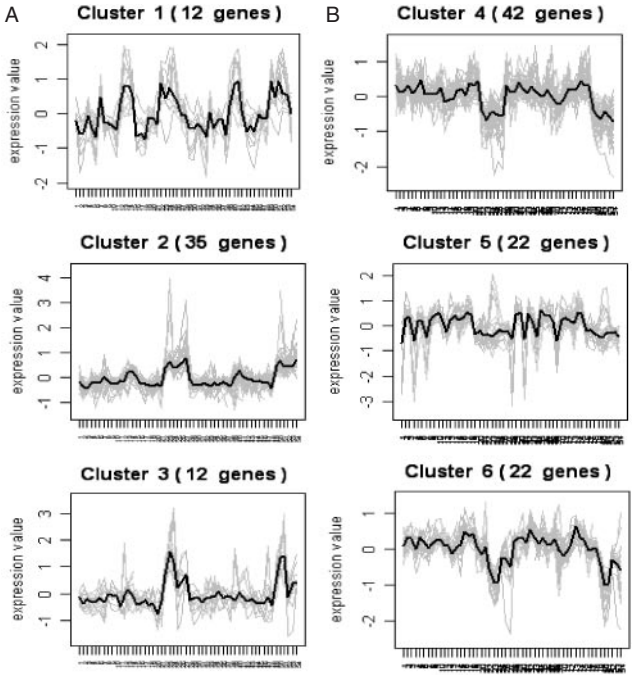

Fig. 1. Data visualization by cluster analysis. The gene expression profile along all 54 experimental conditions (see Table 4 for array labelling) is displayed. (**A**) Genes with a positive $D_{\mathrm{HI}} \times T$ coefficient (induced). (**B**) Genes with a negative $D_{\mathrm{HI}} \times T$ coefficient (repressed). Average expression profile is showed (black line) together with the expression profiles of the genes in the cluster (grey lines).

Applying maSigPro to these data a total of 155 significant genes were selected at a FDR = 0.01 and R-squared threshold equal to 0.6. The FDR gives the expected number of false positives among the selected genes, in this case 1.5, and the R-squared criterion selects for genes which are statistically well modelled. All these 155 genes showed statistically significant differences in the comparison between the high dose and the CO reference group. Out of these, 28 and 91 genes showed also significant differences at the low and medium dose, with respect to the reference CO group, respectively.

Visualization of these expression profiles differences can be performed through the clustering and plotting functions available in the package. A more directed visualization to specific gene expression behaviours is also possible using the values of the estimated regression coefficients. For example, we identified genes having either an induction or repression response upon the HI treatment (with respect to the reference group) by selecting genes with positive or negative values on the estimation of β_9 regression parameter, respectively. This variable gives the slope difference between HI and CO groups when variable $D_{\mathrm{HI}} \times T^2$ is not significant, or the slope difference at Time = 0 between these two groups when the quadratic term is significant. Thus, we obtained 59 genes grouped in the 'induction response' and 86 genes in the 'repression response'. Each of these groups was then subjected to clustering for visualizing the differences between experimental groups (Figs 1 and 2). Figure 1 shows the experiment-wide gene expression profiles, whereas Figure 2 gives the mean profile by groups of each cluster. Figure 1 is useful to evaluate the homogeneity of

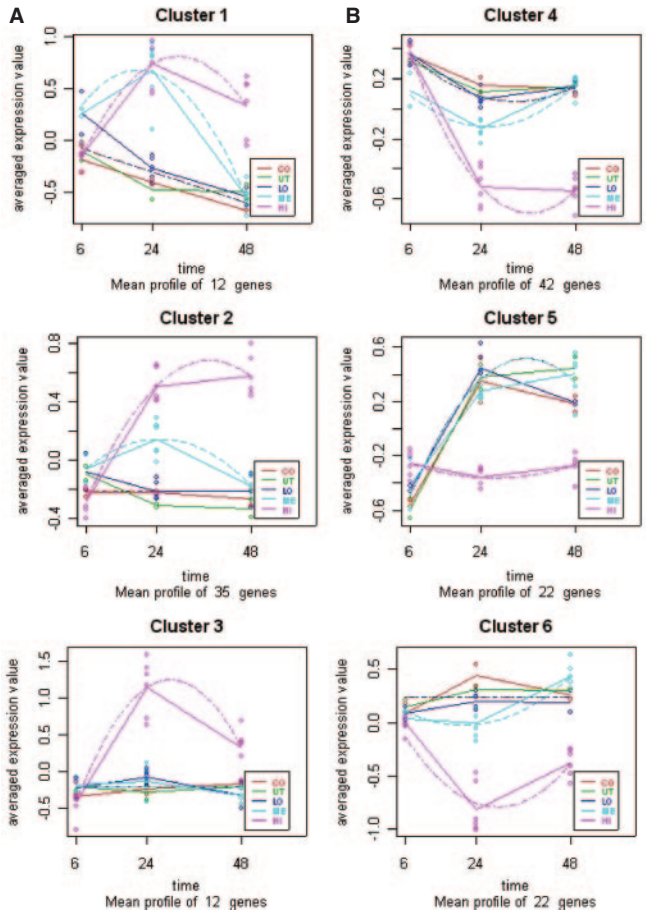

Fig. 2. Data visualization by cluster analysis. Each plot shows the cluster average expression profile by experimental group. (**A**) Genes with a positive $D_{HI} \times T$ coefficient (induced). (**B**) Genes with a negative $D_{HI} \times T$ coefficient (repressed). Dots show actual expression values. Solid lines have been drawn joining the average value of gene expression at each time point for each experimental group. Fitted curves are displayed as dotted lines.

the obtained clusters but the actual profile differences between experimental groups can be better analyzed in Figure 2.

Functional classification of the significant genes showed a high proportion of genes involved in functions related to a toxicological response. Cluster 1 contained a high number of genes related to drug-response, while clusters 2 and 3 were populated by genes involved in protein synthesis, and degradation and maintenance of cell structure. Among the down-regulated genes, many were participating in acute-phase, fatty acid metabolism or had oxidative properties. Interestingly, cluster 4 contained many retinol-signalling and tumorogenesis genes, most of them not found in the original paper analysis (Heijne *et al*., 2003). Overall, maSigPro detected 104 new genes that showed statistically significant differences between experimental groups compared with the result by Heijne *et al*. (2003). These authors used a two-tailed Student's *t*-test per gene on the comparison BB treated (joining LO, ME and HI experimental groups) and CO control with no FDR correction.

To further evaluate the performance of maSigPro we compared our results with those generated using the R package LIMMA. We chose LIMMA for this comparison for being a widely used

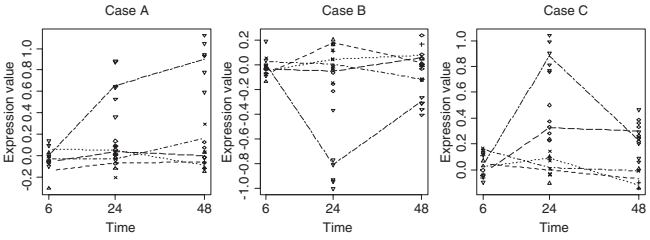

Fig. 3. Three simulated data examples. Different points correspond to the data at different groups. Different lines join mean expression values at each Group × Time combination for the five different groups.

methodology for the statistical analysis of microarray experiments. LIMMA performs a linear fit of the data on the experimental variables and allows setting multiple contrasts for the comparison of the experimental conditions. When applying LIMMA to the bromobenzene study, it became immediately notorious the high number of pairwise comparisons that had to be set to mimic the maSigPro analysis. We focussed on the analysis of gene expression differences between the High dose and the CO group. This implied to analyze the contrasts HI_6h. versus CO_6h., (HI_24h.–HI_6h.) versus (CO_24h.–CO_6h.), (HI_48h.–HI_24) versus (CO_48–CO_24h.) and (HI_48h.–HI_6) versus (CO_48–CO_6h.), and gather the results in one unique gene list. Using this approach, LIMMA selected 63 significant genes at an FDR of 0.01 while maSigPro detected 155. A total of 53 genes were selected by both the methodologies, 10 additional genes were called significant by the LIMMA approach and 102 were solely found by maSigPro. LIMMA exclusive genes showed a greater data variability than those selected with maSigPro. These genes were actually found significant also by maSigPro at the first regression fit but had low *R*-squared values and were consequently not selected. On the other hand, genes detected with maSigPro and not with LIMMA show clear differences between the high doses and corn oil groups. The reasons for the different detection might be attributed to the different criterion for significance between maSigPro and LIMMA. LIMMA applies FDR on the estimated coefficients while maSigPro controls false positives on the significance of each gene model.

3.2 Case 2: simulated data

Since 'live' experimental data cannot tell which genes are truly differentially expressed, we evaluated the power detection of maSigPro on a simulated dataset resembling the structure of the bromobenzene experiment. The dataset contained 600 genes with profile differences which could be classified into 3 expression patterns; single group continuous induction (A), single group transitory repression (B) and differential multi-group induction (C) (see examples of these situations in Fig. 3). In addition, there were 2000 flat profile genes without differences between experimental groups, making a total of 2600 genes. The replicates for each gene were produced as independent observations from a distribution $N(\mu_{ijn}, \sigma^2_{ijn}), i = 1, \ldots, 5;\ j = 1, 2, 3;\ n = 1, \ldots, 2600$. The data were generated considering higher variance to the cases with high gene expression and introducing outliers. We performed 100 independent simulations and computed the number of false positives detected with both maSigPro and LIMMA using an FDR of 0.05. Our results show that maSigPro was successful in controlling the

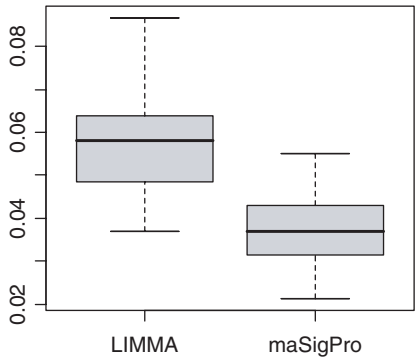

Fig. 4. Results on simulated microarray data. Box-plots summarizing FDR obtained applying LIMMA and maSigPro to 100 simulated datasets.

number of false positive at the given FDR, while LIMMA exceeded this threshold in many cases (Fig. 4). The difference between the FDR obtained with LIMMA and maSigPro is statistically significant at 95% confidence level (0.02 ± 0.002). Furthermore, no type-II errors (false negatives) were present within maSigPro solutions whereas 16% of the simulations analyzed by LIMMA did contain at least one false negative. Further analysis of maSigPro estimates showed that all the significant effects were included in the models and there were $\sim$2.8% of the significant genes with some additional variable in the model, which indicates an adequate control of the false positives at this step of the analysis.

4 DISCUSSION

In this work we present a statistical procedure to identify genes that have different expression profiles among experimental groups in microarray time-course experiments. The method is a two-step regression approach where experimental groups are defined by dummy variables. The first regression fit adjusts a global model and serves to select differentially expressed genes, while in the second step a variable selection strategy is applied to identify statistically significant profile differences between experimental groups. The way variables are defined in the model provides a versatile procedure for studying specific pattern differences among experimental groups and genes.

The choice of using a two-regression steps approach instead of fitting a unique model had a number of reasons and consequences. In principle, it is possible that a model including all the available variables would be statistically significant but would not have any statistically significant coefficient. This situation is possible in multicollinear scenarios. Therefore, it appears more adequate to apply a variable selection strategy to obtain gene-specific models containing only significant variables and where correlated variables had been removed. However, this way of building the models is not very recommendable for the purpose of selecting significant genes. First, because the time necessary to obtain models by steps is much longer than the time needed to estimate a unique model. With datasets including thousands of genes this can become highly time consuming and practically unfeasible. Consequently, it appears much more effective to first fit a global model for all genes, use the ANOVA *P*-values of these global models to find significant genes and then apply stepwise variable selection fit to only this selection of genes. The second reason is based on the outcome of some studies that have shown that regression models created by stepwise approaches yield *P*-values biased towards low values (Harrell, 2002). These *P*-values do not have a proper meaning and their appropriate correction is still a problem. We checked how this circumstance would affect gene expression analysis by applying both the maSigPro approach and solely stepwise regression, with their corresponding *P*-value corrections for multiple comparisons, on different datasets. This experiment showed that Harrell's assertion was true when the goodness of fit of the models (*R*-squared) was not considered, but as the *R*-squared values of the estimated models increased, normally >0.5, both approaches converged. When gene selection uses a high *R*-squared threshold (e.g. 0.6 as used in this example), both approaches yield similar results, but the two-step procedure is computationally less intensive.

Gene selection based on the goodness of fit criterion (high *R*-squared) provides the possibility of selecting genes for which good models could be obtained. This can be in many cases a very interesting option when the researcher is mainly interested in finding biologically meaningful expression trends and in detecting evenly meaningful profiles differences. In this case, high *R*-squared gene models might be successful in capturing these behaviours. In other cases, the aim of the analysis may be the detection of any possible gene expression difference and low *R*-squared models showing some significant coefficients could be allowed. The knowledge of the researcher and the objectives of study in each experiment will help to take a decision about the *R*-squared threshold to use.

Regression approaches rely on a number of assumptions such as independence of the observations, homoscedasticity and normality. Since microarray data might not always meet these requirements, validation of the models would be pertinent. In the simulation study maSigPro successfully detected the existing gene expression profile differences despites the heteroscedasticity and influential values present in the data, indicating that the method is valid for the detection of profile differences in such cases. The maSigPro package provides a series of tools for evaluating the presence of influential data, which is given as one of the results of the analysis process.

In the toxicogenomics example analyzed in this paper, observations were independent because each rat was removed from the experiment after RNA extraction and therefore the measurements had been obtained from different individuals. However, in experiments where gene expression is measured over time on the same subjects the assumption of independence of the observations will not be satisfied. In these cases it would be more recommendable to analyze the data via repeated measures or longitudinal studies (Vittinghoff *et al*., 2005).

Although we have presented the method with $(J-1)$-th time order effects, in experiments where simple gene expression responses are sought or expected and a reduced number of time points are evaluated (<6), quadratic or cubic models would usually be sufficient to analyze the data (note that the polynomial degree is always a maximum, the variable selection step will create in the end models that 'best' fit the data). As already discussed above, it is likely that the researcher is mostly interested in genes which follow biologically meaningful patterns like induction/repression, saturation kinetics or transitory responses, which can easily be

modelled with low degree polynomials. In experiments expanding a larger number of time points, more complex expression patterns could be expected. In this case, simple polynomial models may fail to capture the evolution of gene expression. For such scenarios a piecewise regression or splines regression approach could be applied (Marsh and Cormier, 2001). The inclusion of a splines regression alternative within the maSigPro approach is in principle quite straightforward, as it would simply imply to introduce new dummy variables to define time intervals. The feasibility of this strategy will be addressed in future studies.

The results presented in this work show that maSigPro is a powerful method for the analysis of time-course microarray data. The method detects significant profile differences without carrying out tedious multiple pairwise comparisons, allowing for unbalanced designs and heterogeneous sampling times. The variable definition of the models does permit not only to find genes with temporal expression changes between experimental groups, but also to analyze the magnitude of these differences. The proposed method can easily be extended to include additional variables (e.g. dye) or reduced by removing variables (e.g. to study the evolution over time for one unique group). The availability of the maSigPro methodology as an R package makes this analysis approach easily accessible to the research community.

ACKNOWLEDGEMENTS

The authors thank Wilbert H. M. Heijne from TNO Nutrition and Food Research from Zeist, The Netherlands and Francesco Falciani from School of Biosciences of the University of Birmingham, United Kingdom for providing us with the datasets used in the development of maSigPro. Furthermore the authors thank Francesco for his helpful comments.

Conflict of Interest: none declared.

REFERENCES

Bar-Joseph,Z. *et al.* (2003) Comparing the continuous representation of time series expression profiles to identify differentially expressed genes. *Proc. Natl Acad. Sci. USA*, **100**, 10146–10151.

Bar-Joseph,Z. (2004) Analyzing time series gene expression data. *Bioinformatics*, **20**, 2493–2503.

Beal,M.J. *et al.* (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, **21**, 349–356.

Draghici,S. (2003) *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC, London.

Draper,N. and Smith,H. (1998) *Applied Regression Analysis*. 3rd edn. Wiley, New York.

Ernst,J., Nau,G.J. and Bar-Joseph,Z. (2005) Clustering short time series gene expression data. *Bioinformatics*, **21** (Suppl.1), 159–168.

Harrell,F. (2002) *Regression Modeling Strategies: With Applications To Linear Models, Logistic Regression And Survival Analysis*. Springer, New York.

Heijne,W.H.M. *et al.* (2003) Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach. *Biochem. Pharmacol.*, **65**, 857–875.

Herrero,J. *et al.* (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

Kerr,M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Liu,H. *et al.* (2005) Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments. *BMC Bioinformatics*, **6**, 106.

Luan,Y. and Li,C. (2003) Clustering of time-course gene expression data using a mixed-effects models with B-splines. *Bioinformatics*, **19**, 474–482.

Lukashin,A.V. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

Marsh,L.C. and Cormier,D.R. (2001) Spline regression models. Sage Publications, Inc, Thousand Oaks, California.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Park,T. *et al.* (2003) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, **19**, 694–703.

Peddada,S.D. *et al.* (2003) Gene selection and clustering for time-course and dose–response microarray experiments using order-restricted inference. *Bioinformatics*, **19**, 834–841.

Reiner,A. *et al.* (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, article 3.

Speed,T. (2003) Statistical Analysis of Gene Expression Microarray Data. Chapman & Hall/CRC, London.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9** (12), 3273–3297.

Tusher,V. *et al.* (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl Acad. Sci.*, **98**, 5116–5121.

Vittinghoff,E. *et al.* (2005) *Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models*. Springer, New York.

Wolfinger,R.D. *et al.* (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

Xu,X.L. *et al.* (2002) A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Hum Mol Genet.*, **11**, 1977–1985.