BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 14 2007, pages 1792–1800
doi:10.1093/bioinformatics/btm251

Gene expression

Discovering gene expression patterns in time course microarray experiments by ANOVA–SCA

María José Nueda[1,*,†], Ana Conesa[2,6,†], Johan A. Westerhuis[3], Huub C. J. Hoefsloot[3], Age K. Smilde[3,4], Manuel Talón[2] and Alberto Ferrer[5]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Alicante, Apartado 03080, Alicante, [2]Centro de Genómica, Instituto Valenciano de Investigaciones Agrarias, Apartado Oficial 46113, Moncada, Spain, [3]Biosystems Data Analysis, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV, Amsterdam, [4]TNO Quality of life, PO Box 360 AJ Zeist, The Netherlands, [5]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universidad Politécnica de Valencia, Cno. Vera s/n, Edificio I-3, Apartado 46022 and [6]Bioinformatics Department, Centro de Investigación Príncipe Felipe, Autopista del Saler, 16, E46013, Valencia, Spain

Received on February 15, 2007; revised on April 17, 2007; accepted on May 2, 2007

Advance Access publication May 22, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Designed microarray experiments are used to investigate the effects that controlled experimental factors have on gene expression and learn about the transcriptional responses associated with external variables. In these datasets, signals of interest coexist with varying sources of unwanted noise in a framework of (co)relation among the measured variables and with the different levels of the studied factors. Discovering experimentally relevant transcriptional changes require methodologies that take all these elements into account.

Results: In this work, we develop the application of the Analysis of variance–simultaneous component analysis (ANOVA–SCA) Smilde *et al*, *Bioinformatics*, (2005) to the analysis of multiple series time course microarray data as an example of multifactorial gene expression profiling experiments. We denoted this implementation as *ASCA-genes*. We show how the combination of ANOVA-modeling and a dimension reduction technique is effective in extracting targeted signals from data by-passing structural noise. The methodology is valuable for identifying main and secondary responses associated with the experimental factors and spotting relevant experimental conditions. We additionally propose a novel approach for gene selection in the context of the relation of individual transcriptional patterns to global gene expression signals. We demonstrate the methodology on both real and synthetic datasets.

Availability: ASCA-genes has been implemented in the statistical language R and is available at http://www.ivia.es/centrodegenomica/bioinformatics.htm.

Contact: mj.nueda@ua.es and aconesa@cipf.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Multiple series time- course (MSTC) microarray experiments are designed experimental setups in which gene expression is measured at various points of a given time interval on samples that correspond to different levels of other experimental factor(s), such as treatment, tissue or strain. As in many other functional genomics datasets, MSTC data contain information about a large number of variables (genes) measured on a relatively small number of samples (experimental conditions). The analysis of this kind of data is usually addressed either as the identification of co-expressing genes clusters, or as the detection of differentially expressed genes. Traditional clustering methods have been applied to the analysis of microarray time course data (Lukashin and Fuchs, 2001; Spellman *et al*., 1998) and more recently dedicated clustering algorithms have been developed that particularly consider the temporal property of gene expression (Bar-Joseph *et al*., 2003). These approaches are efficient in finding groups of co-expressing genes but when the experimental setup is complex (different numbers of treatments, replicates, dye-swaps, etc.) the evaluation of the results on the basis of the clustered patterns can become a rather complicated task. Furthermore, clustering methods tend to equally weight all samples while deriving gene partition, which could not be the most convenient approach when expression changes are only present in a subset of conditions. A second type of methodologies aims at the identification of genes whose expression vary across experimental conditions in a statistically significant manner (Conesa *et al*., 2006; Storey *et al*., 2005; Tai and Speed, 2004). These approaches are frequently univariate and as such do not provide the adequate framework for generating a global understanding of the information contained in the data.

In general, when managing large amounts of noisy but correlated data, such as in the case of microarray experiments and especially when various experimental factors and levels combine, data analysis can greatly benefit from approaches that generate information about major and secondary patterns

*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.

of variability present through the experimental setup. Such explorative approaches are effective in providing a global understanding of the effects that the different factors cause on gene expression, help in identifying most relevant experimental conditions and can shed light on how to address subsequent statistical analysis, e.g. which would be the contrasts of greatest interest. Dimensionality reduction techniques, such as principal component analysis (PCA) are suited for explorative and summarizing analyses in these datasets as they are able to model the relationships between genes by analyzing the correlation structure of the data (Quakenbush, 2001). PCA in microarray data was introduced by Raychadhuri *et al.* (2000) for the analysis of Chu's yeast sporulation dataset (Chu *et al.*, 1998). These authors showed the basis of principal component (PC) interpretation in the gene expression framework indicating the possibilities and difficulties of using PCA as a clustering technique. Other studies have applied PCA and related dimension reduction techniques in microarray data analysis for the purpose of classifying samples (Dai *et al.*, 2006; Landgrebe *et al.*, 2002; Nguyen and Rocke, 2002), finding co-expressing genes (Yeung and Ruzzo, 2001) or identifying odd data (Hilsenbeck *et al.*, 1999). Methods have been developed to introduce statistical significance in the choice of PCs or for selecting relevant genes (Landgrebe *et al.*, 2002; Roden *et al.*, 2006). However, these approaches generally do not take the underlying experimental design into account. Therefore, the different sources of variation are confounded in the PCA model and this can seriously hamper the interpretation of the principal components (Jansen *et al.*, 2005). The typical methodology for the analysis of designed experiments is Analysis of variance (ANOVA), which focuses on the separation of the different sources of variability. Smilde and co-workers (Smilde *et al.*, 2005) proposed an adaptation of the SCA (simultaneous component analysis) algorithm that incorporates experimental design information through ANOVA modeling. The so-called ASCA (ANOVA–SCA) basically applies PCA to the estimated parameters in each source of variation of an ANOVA model. This methodology has been successfully applied to data analysis problems in psychology (Timmerman and Kiers, 2003) and metabolomics (Jansen *et al.*, 2005).

In this work, we study further the applicability of the ASCA approach to the analysis of high-dimensional microarray data from a designed experiment involving several factors. In particular, we use ASCA to explore gene expression trends and differences in MSTC microarray experiments. We show how ASCA is an effective approach for separating the data variability present in a complex MSTC experimental setup to (i) extract the signal of interest from noisy data, (ii) reveal major expression patterns associated to the different experimental factors and (iii) identify most relevant experimental conditions. We develop further the original methodology by incorporating algorithms for identifying significant signals and selecting genes that behave according to the detected patterns. We use a published MSTC dataset for illustrating the methods presented in this work and synthetic data to provide a deeper understanding of the working of the methodology. The methodology, denoted by ASCA-genes, has been implemented in the statistical language R and is available from the authors.

2 METHODS

2.1. Model definition

We will consider the general case of a MSTC microarray experiment, where the experimental design is defined by two main types of variables: the time component and the experimental groups for which temporal gene expression differences are sought. Consider I time points ($i=1,\ldots,I$), J experimental groups ($j=1,\ldots,J$), R_{ij} replications, $r=1,\ldots,R_{ij}$ for each case ij and N genes ($n=1,\ldots,N$). For each gene, we will denote by x_{ijr} the gene expression measure at the time i, under condition j and for replicate r.

The analysis of this experiment with the ASCA approach (Smilde *et al.*, 2005) implies the definition for each gene of the ANOVA model given in Equation (1)

$$x_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + (\alpha\beta\gamma)_{ijr} \quad (1)$$

where, μ is an offset term, α_i is the model parameter for factor time on level i, β_j measures the j-th group effect, $(\alpha\beta)_{ij}$ represents the interaction effect between the i-th time and j-th group, and the individual variation is indicated by $(\alpha\beta\gamma)_{\mathrm{ijr}}$ instead of $\varepsilon_{\mathrm{ijr}}$ to avoid confusion with the error term in the subsequently derived ASCA model.

When we consider a microarray experiment with N genes and $M=\sum_{i,j} R_{ij}$ samples, a matrix $\boldsymbol{X}$ of dimensions $(M \times N)$ can be defined containing the entire gene expression dataset. Similarly, the estimates of the ANOVA parameters on the right hand side of Equation (1) can be obtained for all genes and collected into matrices, where rows represent samples and columns represent genes. Therefore expression (2) can be obtained:

$$X = 1\boldsymbol{m}^T + \boldsymbol{X}_{\mathrm{a}} + \boldsymbol{X}_{\mathrm{b}} + \boldsymbol{X}_{\mathrm{ab}} + \boldsymbol{X}_{\mathrm{abg}} \quad (2)$$

where, 1 is a size M column vector of ones, $\boldsymbol{m}^T$ is a size N row vector containing estimates of μ for each gene, matrices $\boldsymbol{X}_a$, $\boldsymbol{X}_b$ and $\boldsymbol{X}_{ab}$ contain the estimates of parameters α_i, β_j and $(\alpha\beta)_{ij}$, respectively, and $\boldsymbol{X}_{abg}$ contains the residuals named $(\alpha\beta\gamma)_{ijr}$. The rows of matrices $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ are highly structured. All rows related to one level i of factor α are equal in $\boldsymbol{X}_a$ and analogously all rows of $\boldsymbol{X}_b$ are equal for each level j of factor β (see Jansen *et al.*, 2005 for details).

Applying multivariate projection techniques in matrices $\boldsymbol{X}_a$, $\boldsymbol{X}_b$, $\boldsymbol{X}_{ab}$ and $\boldsymbol{X}_{abg}$ the information can be summarized. Consequently, the ASCA-model corresponding to ANOVA Equation (2) is given in Equation (3),

$$X = 1\boldsymbol{m}^T + \boldsymbol{T}_a\boldsymbol{P}_a^T + \boldsymbol{T}_b\boldsymbol{P}_b^T + \boldsymbol{T}_{ab}\boldsymbol{P}_{ab}^T + \boldsymbol{T}_{abg}\boldsymbol{P}_{abg}^T + \boldsymbol{E} \quad (3)$$

where, the SCA component scores of each submodel are given by the matrices indicated by $\boldsymbol{T}_a$, $\boldsymbol{T}_b$, $\boldsymbol{T}_{ab}$, $\boldsymbol{T}_{abg}$, and the submodel loadings are given by the matrices $\boldsymbol{P}_a$, $\boldsymbol{P}_b$, $\boldsymbol{P}_{ab}$, $\boldsymbol{P}_{abg}$, where $\boldsymbol{P}_x^T\boldsymbol{P}_x = I$ for $x = a$, b, ab or abg, without loss of generality. $\boldsymbol{E}$ is a matrix in which the residuals of all submodels of the ASCA-model are collected: $\boldsymbol{E}=\boldsymbol{E}_a+\boldsymbol{E}_b+\boldsymbol{E}_{ab}+\boldsymbol{E}_{abg}$.

In the rest of this work, the ASCA submodels in Equation (3) will be indicated as 'submodel a', 'submodel b', 'submodel ab' and 'submodel abg', respectively. Further details on the ASCA approach can be found in Jansen *et al.* (2005).

Once the ASCA-model has been derived and computed, data analysis proceeds, as in regular PCA, with the exploration of the loadings in a selected number of PCs, which are typically obtained on the basis of the percentage of the explained variability or by a cross-validation criterion. In the case of ASCA, the number of components to retain has to be decided for each of the submodels. We propose a PC selection procedure based on the screeplot of each submodel and the explained variability and interpretation of the associated patterns. Main PCs are identified as those previous to the slope inflexion on a scree-plot representation of the accumulated variability and additional PCs are retained when they describe interesting expression patterns. Finally, the

graphical analysis of the score profiles of the selected components in the different submodels (time, experimental groups and interactions, $\boldsymbol{T}_a$, $\boldsymbol{T}_b$, $\boldsymbol{T}_{ab}$, respectively) allows to extract conclusions on the effects that the different experimental factors have on gene expression.

2.2 Gene-wise analysis

Once major variability patterns have been identified, one step further in the analysis is to identify both those genes that more closely follow the detected trends as well as those that clearly diverge from the general model. The first ones would represent those genes that most co-coordinately respond in the experimental context; and the second ones would include odd behaviors or outlier data. For this goal we analyze the loadings ($\boldsymbol{P}_a$, $\boldsymbol{P}_b$, $\boldsymbol{P}_{ab}$) and the residuals ($\boldsymbol{E}_a$, $\boldsymbol{E}_b$, $\boldsymbol{E}_{ab}$) of the genes in the components selected in each model. Genes with high absolute loading in a specific component are those that follow the behavior described by this component and genes with high residuals are genes poorly modeled by this component. We propose the use of two statistics, the leverage and the squared prediction error (SPE) to quantify these two aspects.

The leverage is a measure of the importance of a variable (in this case gene) in the PCA model. This is computed according to Equation (4) (Martens and Næs, 1989):

$$\boldsymbol{h}_x = \text{diag}[\boldsymbol{P}_x\boldsymbol{P}_x^T], x = a, b \text{ or } ab \quad (4)$$

Where $\text{diag}[\boldsymbol{P}_x^T\boldsymbol{P}_x]$ is a vector with the diagonal of matrix $\boldsymbol{P}_x^T\boldsymbol{P}_x$, and $\boldsymbol{h}_x$ the vector containing the leverage values for all the genes in submodel x ($x = a$, b or ab). A threshold for statistical significance of leverage can be defined by resampling methods. In our case, we have chosen a permutation approach in which a number of row permutations of matrix $\boldsymbol{X}$ are generated to create a reference distribution where the designed structure of the data has been destroyed. ASCA is then applied to each permuted matrix with the same number of components as taken for the original data and gene leverages are computed in each case. The leverage threshold value at a given confidence $(1-\alpha)$ is obtained as the average of the $(1-\alpha)$% quantiles computed for all the genes.

The SPE associated to a particular gene is a measure of the goodness of fit of the model for that specific gene. Genes not following the general structure defined in the fitted model will have high SPE. The vector containing the values of this statistic for all the genes can be computed in each submodel according to Equation (5):

$$\boldsymbol{SPE}_x = \text{diag}[\boldsymbol{E}_x^T\boldsymbol{E}_x], x = a, b \text{ or } ab \quad (5)$$

The SPE for a particular gene in a submodel is a quadratic form of the errors associated with that gene. Assuming that these errors are well approximated by a multivariate normal distribution, Box (1954) showed that the SPE is well approximated by a weighted χ^2-distribution ($g\chi_h^2$). We have used this approximation to establish the $(1-\alpha)$ confidence SPE threshold. We estimate g and h by matching moments of the $g\chi_h^2$ distribution: the mean and variance ($\mu = gh$, $\sigma^2 = 2g^2h$) are equated to the sample mean (m) and variance (v) of the SPE sample obtaining the next expression as SPE threshold at α level of significance:

$$SPE_\alpha = \frac{v}{2m}\chi^2_{(2m^2/v),\alpha} \quad (6)$$

Therefore, by combining leverage and SPE criteria allows genes can be categorized in relation to modeling and interest. Most relevant genes in the derived ASCA-model will be those showing high leverage and low squared prediction error. Poorly modeled genes will be identified by high values in their SPE, while those genes having low leverage and low SPE will be regarded as not affected by the experimental factors (Table 1).

Table 1. Criteria for gene categorization. Shaded categories provide genes for further analysis

	Low SPE	High SPE
Low leverage	Not responsive	Badly modeled Possibly odd data
High leverage	Well modeled Follow main trends	Influential but poorly modeled

2.3 Comparative analysis

The ASCA-genes approach was compared with four different methodologies described for the analysis of time course microarray data: the clustering methods SOTA and *K*-means and the hypothesis testing based approaches timecourse and maSigPro. SOTA is a hierarchical unsupervised growing neural network which adopts the topology of a binary tree and offers a statistical criterion for cluster division (Herrero *et al*., 2001). *K*-means is a non-hierarchical partition based algorithm widely used in microarray data analysis (Hartigan and Wong, 1979). *K*-means uses a minimum 'within-class sum of squares from the centers' criterion to select the clusters and requires the number of partitions to be fixed in advanced. Timecourse applies a multivariate empirical Bayes statistic (the *MB*-statistic) to the analysis of replicated time course data. The algorithm contrasts, for each gene, the null hypothesis of constant vector of means along the time component, to the alternative hypothesis of non invariability. The *MB*-statistic can be used to rank genes in the order of evidence of non-constancy (Tai and Speed, 2004). Finally, maSigPro is a model-based univariate method in which different temporal series are modeled by binary variables. The method assesses significant differences in gene expression profiles between time series through the significance of the estimated model parameters (Conesa *et al*., 2006).

SOTA analysis of time course data was done at the GEPAS server (www.gepas.org) taking Euclidean distance as similarity metric and a threshold of 90% node variability as stop growing tree condition. *K*-means, timecourse and maSigPro analyses were performed with the corresponding R packages available at the Bioconductor repository (www.bioconductor.org). The default Hartigan and Wong algorithm and 25 partitions were taken as parameters of the *K-means* function. A criterion of positive *MB*-statistics was used for the feature selection in the time course package while a significance value of 0.01 was applied in the maSigPro approach. Additionally, PCA computations were done using the *princomp* function of the stats R package.

3 RESULTS

The proposed method has been applied to two datasets. The first one consists of a toxicogenomics study involving different treatments and time points. The second is a synthetic dataset which reproduces the structure present in the toxicogenomics experiment. In the latter dataset, signals and noise sources have been simulated to resemble real data. The synthetic data was used to analyze how different sources of variability are treated by the ASCA-genes approach while the real dataset was used to study the biological interpretation of the ASCA-genes results.

The experimental dataset comes from a toxicogenomics study by Heijne *et al*. (2003) where the effect of the hepatotoxicant bromobenzene is studied. In this study, groups of rats were treated with three different doses of this toxic compound dissolved in corn oil. There were two additional groups of rats

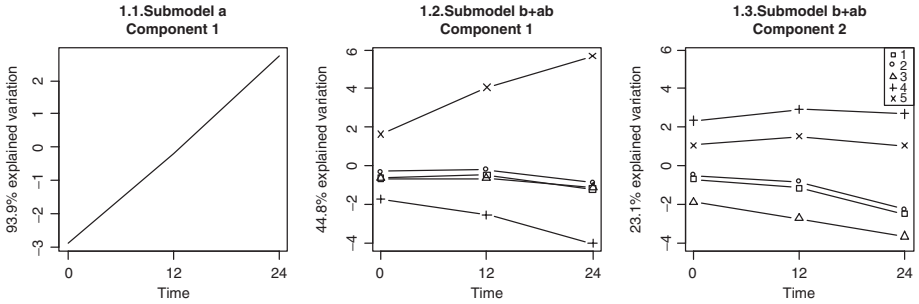

Fig. 1. Score profiles: (**1.1**) Component 1 in 'submodel a', (**1.2**) Component 1 in 'submodel b+ab' and (**1.3**) Component 2 in 'submodel b+ab'.

without toxic treatment: an untreated group and a group where only the drug administration vehicle was given. In total there were five experimental groups denoted by the labels UT (untreated), CO (corn oil), LO (low dose), ME (medium dose) and HI (high dose). One to three rats per experimental group were sacrificed at three different time-points (6, 24 and 48 h) to extract liver RNA. Individual rat RNA samples were co-hybridized against an external reference and hybridizations were duplicated swapping the two labeling dyes. In total there were 54 different hybridizations and 2665 genes available for statistical analysis. For further details on this dataset see Heijne *et al*. (2003). Therefore, this example contains three time points ($i=1,2,3$), 5 experimental groups ($j=1,\ldots,5$), 2 or 6 replicates, $r=1,\ldots,R_{ij}$ (2 or 6) for each case ij and 2665 genes ($n=1,\ldots,N$).

The simulated data was created to reproduce the experimental set up of the previous dataset—five experimental groups and three time points—introducing responsive genes and noise in a controlled manner. The synthetic dataset contained 410 genes with profile changes classified into five expression patterns: 100 genes with continuous induction for all the groups (A), 100 genes with continuous induction for group 5 (B), 100 genes with continuous repression for group 5 (C), 100 genes with continuous induction for group 4 (D) and only 10 with transitory induction for group 3 (E). The reason for including a responsive group with few genes, group (E), is the interest in analyzing the behavior of the method on a minor trend. Additionally, there were 2500 flat profile genes without differences between experimental groups, making a total of 2910 genes. The replicates for each gene were produced as independent observations from a normal distribution. It is however, known that in real experiments residuals do not necessarily follow a normal distribution. Non-random sources of technical variability such as spatial bias, not corrected dye-swap labeling, or mixture of data from different labs or experimenters can deviate data from Gaussian distributions. Therefore, to better reproduce real microarray data and in order to analyze how ASCA-genes behaves with this type of variability, systematic noise was introduced to the dataset by splitting the data set in two replicates and adding two opposite normal distributions to each half.

As the purpose of this study is to identify gene expression profile differences between experimental groups, when applying ASCA-genes the choice is made to join for each gene β_j and $(\alpha\beta)_{ij}$ effects in Equation (1) and analyze them in one submodel as it is shown in Equation (7) (Jansen *et al*., 2005).

$$x_{ijr}=\mu+\alpha_i+[\beta_j+(\alpha\beta)_{ij}]+(\alpha\beta\gamma)_{ijr}=\mu+\alpha_i+(\beta+\alpha\beta)_{ij}+(\alpha\beta\gamma)_{ijr} \quad (7)$$

Out of these effects, $(\beta+\alpha\beta)_{ij}$ is the most important for biological interpretations: it represents the effect of the treatment group on the gene expression measured as deviations from the common time effect α_i for each gene. The estimates of the ANOVA parameters in Equation (7) can be arranged for all the genes into matrices obtaining Equation (8),

$$X=\mathbf{1}\boldsymbol{m}^T+\boldsymbol{X}_a+\boldsymbol{X}_b+\boldsymbol{X}_{ab}+\boldsymbol{X}_{abg}=\mathbf{1}\boldsymbol{m}^T+\boldsymbol{X}_a+\boldsymbol{X}_{b+ab}+\boldsymbol{X}_{abg} \quad (8)$$

where 'submodel a' describes the variation due to the factor time, 'submodel b+ab' describes all the variation related to the factor treatment group, and 'submodel abg' describes the residual (random and non-random) variation in the data. Therefore, the ASCA-model applied to the described datasets is given by Equation (9).

$$X=\mathbf{1}\boldsymbol{m}^T+\boldsymbol{T}_a\boldsymbol{P}_a^T+\boldsymbol{T}_{b+ab}\boldsymbol{P}_{b+ab}^T+\boldsymbol{T}_{abg}\boldsymbol{P}_{abg}^T+\boldsymbol{E} \quad (9)$$

3.1 Case 1: simulated data

By fitting the ASCA model to the synthetic dataset and using the PC selection criterion described, one and two components were selected for 'submodel a' and 'submodel b+ab', respectively.

The score profiles of the components of the submodels reveal the most common expression patterns in the genes. The first component of the 'submodel a' (Fig. 1.1) shows a positive linear effect through time for all experimental groups. This pattern is the case to the simulated pattern A, where no time-experimental group interactions were modeled, only a time effect in all the groups. The first component of the 'submodel b+ab' (Fig. 1.2) shows different behavior for group 5 and group 4 with respect to the other groups: a clear positive linear effect through time for group 5 (related to patterns B and C) and a light negative one for group 4 (related with pattern D). The second component of the same submodel (Fig. 1.3) shows positive differential behavior of group 4 and group 5 (related to patterns D, B and C) and also signals different behavior of groups 3 (related to pattern E), even this is less pronounced.

By analyzing the loadings of the genes, we can detect genes that follow the trends shown in the score profiles of the components of the ASCA-model. For example, if we focus on

'submodel b+ab' (Fig. 2), genes with a high positive loadings value for the first component of this submodel correspond to genes which were simulated in pattern B, while genes with a high negative value in the same component were genes simulated as pattern C. As pattern D is detected negatively by the first component and positively for the second one, genes from pattern D have negative loadings for the first component and positive loadings for the second. Another interesting result is that genes without interactions between time and groups (simulated pattern A and modeled by 'submodel a') and flat profile genes have low loadings in this model.

The analysis of the leverages and SPE in both models shows interesting results (Fig. 3). First of all, 'submodel a' shows high leverages and good modeling for case A, which is the case where the time effect is the same for all the groups. Secondly, 'submodel b+ab' in general shows high leverages and good modeling for cases B, C and D, low leverages for case A and Flat profiles, and high leverages and high SPE for case E. This indicates that case E is badly represented by the model. The reason for this is presumably due to the fact that the variability associated with case E represents a small percentage of the gene set, and hence this source of variation is not included in the model. This example illustrates that when there is a small group of genes whose behavior is not described in the model, they can

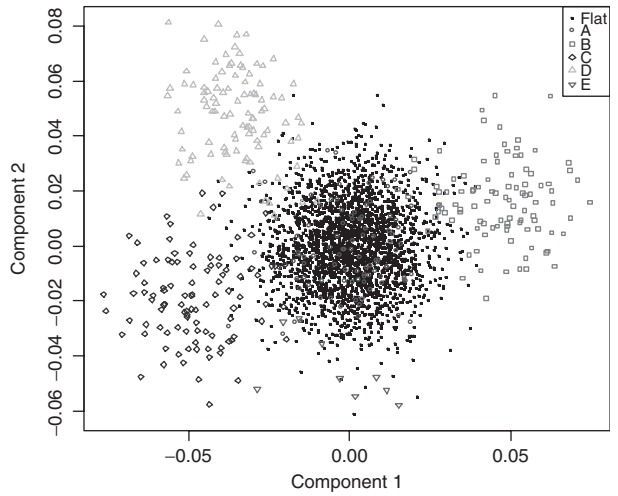

Fig. 2. Gene loadings in the two components of 'submodel b+ab'.

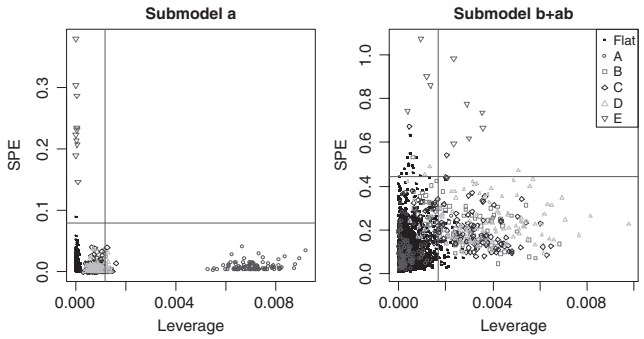

Fig. 3. Leverages and SPE of the genes for the simulated dataset in 'submodel a' and 'b+ab'.

be detected by residuals analysis or in extra components. The thresholds shown in Figure 3 have been computed with 100 permutations and $\alpha = 0.01$ using the gene-wise analysis described previously. Taking into account that genes of interest have either high leverages or high SPE the method points to 426 genes. This selection means a sensitivity (defined as the proportion of detections above true positives) of 91% and a specificity (defined as the ratio between false calls and true negatives) of 98%.

To illustrate how ASCA-genes analysis is able to filter non-random or unwanted sources of noise and extracts the variability associated to the experimental factors we decided to compare the method with a general projection technique: PCA. We applied PCA to this synthetic dataset and analyzed the variability and meaning of the components similarly to the procedure followed with ASCA-genes. Two components were extracted according to the scree-plot criterion. We observe that the first PCA component explains 82.1% of the total variation. However, as it can be seen on the first plot of Figure 4 this component cannot be associated to the behavior of any of the simulated patterns (A, B, C, D or E). When sample scores were analyzed for this first component, we observed that this PC represents the technical noise introduced to the data. The second component (3.94% explained variation) identifies the pattern B and the opposite C. Inspection of a possibly interesting third component (2.6% explained variation) revealed a general linear tendency in all the groups (patterns A) and the pattern simulated in D (data not shown). Although these results are interesting, they pose a difficulty to interpretation as the by far main direction of variability turns out to be associated to a technical feature rather to the experimental factors which are the goal of the study. When the identification of these sources of non-random noise is not possible or at least easy—as could be the case in real data, we would probably fail to give a biological interpretation to the results obtained by this approach. In short, because PCA does not impose the experimental design when doing variability analysis, targeted and non-targeted effects can show up and confound.

This problem is directly addressed by ASCA. By taking into account the experimental design, the methodology focuses on the signals of interest avoiding the interference of non-random noise effects. Table 2 shows the amount of variation associated to each experimental factor and the percentage explained in PCA and ASCA-genes in this example. In PCA, the variation has been computed using ANOVA on the fitted principal components. This is the so-called PC-ANOVA (Jansen *et al.*, 2005). We can observe that while PCA mainly recovers the variability structure present in the residuals ('submodel abg'), the ASCA-genes procedure focuses in the variation of 'submodels a' and 'b+ab' directly.

3.2 Case 2: experimental data

Once the technical aspects of the ASCA-genes procedure has been revealed on a synthetic dataset, we study the methodology on a real experiment to analyze the biological meaning of the results generated by this approach. We have used for this the toxicogenomics study described earlier.

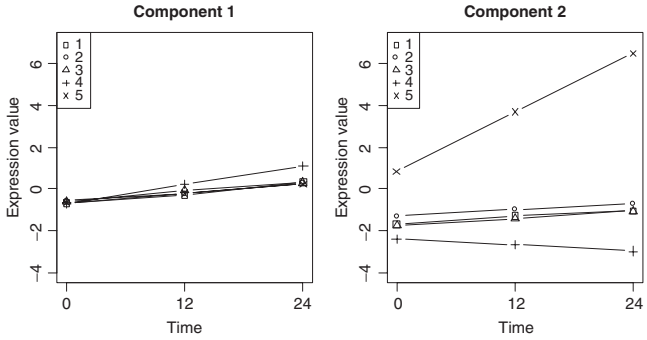

Fig. 4. Score profiles of the two components of PCA model in the simulated dataset. Lines join the averages for each group and time-point.

Table 2. Variation of the submodels explained by PCA and ASCA-genes

		PCA (2PCs)		ASCA-genes (1 and 2 PCs)	
Model	Initial	Explained	%	Explained	%
a (time)	272	32	11.76	255.3	**93.86**
b + ab (group + interaction)	1053	461	43.78	713.9	67.80
abg (residual)	10815	9954	**92.04**	0	0
Total	12140	10447	86.05	969.2	7.98

Values collecting most variability in each approach are given in bold.

Firstly, data exploration by PCA showed that the first component of variability was associated to the dye labeling of the samples, and only on the second PC revealed a distinct behavior for the high bromobenzene dose (Fig. 5). This result illustrates the structural noise problem described in the previous section.

For ASCA analysis of these data, Equation (8) was applied to model the gene expression response. The component selection procedure gave as a result one and three components for 'submodel a' and 'submodel b+ab', respectively. The score profile of the first component of the 'submodel a' (submodel for the time factor, explaining 75.15% of variability) shows that in general, gene expression is mostly affected after 24 h of treatment, followed by a slight reversion at 48 h (Fig. 6). This result reveals the time frame in which treatments have the strongest effect on gene expression and might indicate either a recovery of the biological systems or a loss of drug action at 48 h. The score profiles of the three components of the 'submodel b+ab' (treatment plus interaction submodel) describe the different responses in time (gene expression patterns) for the treatment groups (Fig. 7). These score profiles show a different effect of the bromobenzene doses on gene expression through time. The score profile of the first component identifies a marked effect of the high bromobenzene dose in gene expression which is different to the rest of treatments. As this component represent almost 50% of the

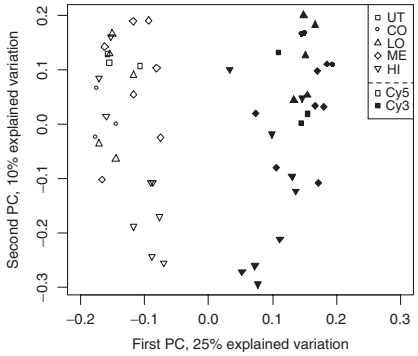

Fig. 5. PCA scores of bromobenzene data.

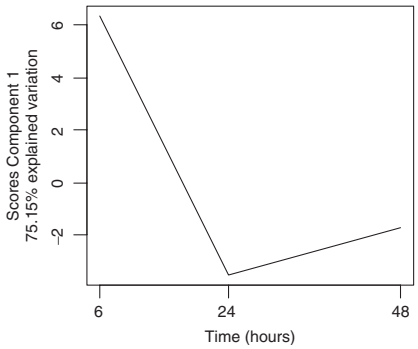

Fig. 6. Score profiles of component 1 of "submodel a".

variability associated to this model, it can be interpreted that high bromobenzene is, by far, the treatment that most affects gene expression. The next two components had a lower weight (they explain around 10% of the data variability each) and thus represent behaviors of lower impact. Second component identifies a gene expression pattern characterized by a difference in the gene expression response between the HI and ME treatments at 24 h, while the third component detects basically an effect of differential expression at 6 h for the HI and ME doses. In general, the ASCA-genes analysis indicates that the major gene expression response of this study is focused on the HI doses at 24 h and there is a lower response to medium bromobenzene doses. Interestingly, similar conclusions about the patterns of bromobenzene toxicity were obtained by Jansen *et al*. (2005) on the analysis of metabolic changes present in the rats used for this study.

The identification of genes following the patterns given by the score profiles of the components is done by loading analysis. Genes with high absolute loading values for the first component of the 'submodel b+ab' are genes for which the bromobenzene produces an effect in the high doses very different to the rest of the doses: in case of positive loading, in the high doses gene expression decreases (repression) at 24 h and it is maintained at 48 h, while the remaining groups do not vary very much over time; on the other hand, genes with negative loadings for this component have the opposite described effect (induction). Finally, genes with loadings on this component near 0 have a pattern different to that described by the first component. Similar reasoning can be applied to the second and third components.

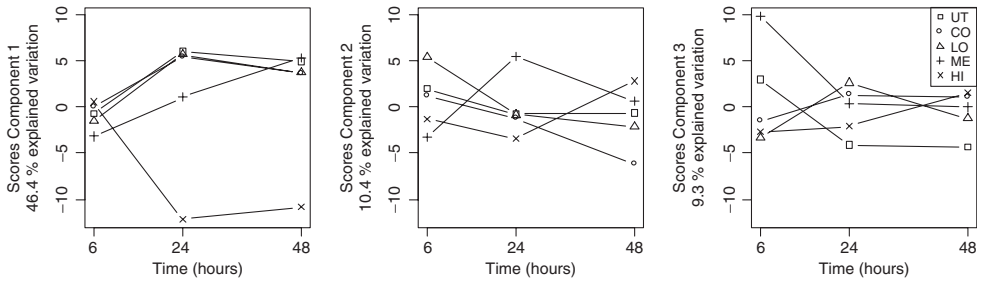

Fig. 7. Score profiles of the first three components of submodel 'b+ab'.

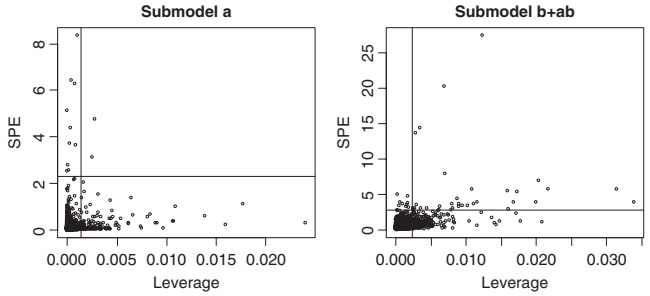

Fig. 8. SPE and leverage statistics of the genes in 'submodel a' (left) and 'submodel b+ab' (right). SPE and leverage cut-off values are indicated by horizontal and vertical lines, respectively.

Genes well modeled by the ASCA-model will have high leverages; whereas poorly modeled genes will show high SPE values. Figure 8 shows SPE and leverage values computed for all the genes in 'submodels a' and 'b+ab'. In general, genes with high leverage exhibit low SPE value, which means that significant model contributions are also well modeled genes. Cut off values for SPE and leverage were computed as described in the methods section taking $\alpha = 0.05$. In total, 345 genes were found with SPE and leverage values beyond their respective thresholds, of which 157 in 'submodel a', 247 in 'submodel b+ab' and 59 in both submodels. Additionally 28 genes were identified as high SPE genes.

To investigate the biological meaning of the gene selection provided by this approach, Gene Ontology (GO) annotations were fetched for the collection of genes present in the rat chip and functional class enrichment analysis was executed using the software Blast2GO (Conesa *et al*., 2005) taking a false discovery rate control level of 0.05. The ASCA-genes selected gene pool was significantly enriched for functional categories such as glutathione transferase activity, oxidoreductase activity, microsome, heme binding, fatty acid metabolism, steroid biosynthesis, xenobiotic metabolism, ferric ion binding, response to stimulus, secondary metabolism, fibrinogen complex and structural component of ribosome. These categories are in agreement with a detoxification response described by Heijne *et al*. (2003) that includes upregulation of redox enzymes, such as microsomal glutathione-S transferase or heme oxygenase, which act in the degradation of xenobiotic compounds, the induction of ribosomal constituents and the regulation of acute-phase related proteins such as ferritin and fibrinogen. Steroid biosynthesis and fatty acid metabolism are pathways reported to be induced by CO administration trial animals (Takashima *et al*., 2006). This result indicates a meaningful biological content in the genes associated to the main transcriptional responses detected by the proposed procedure. Additionally, high SPE genes included epoxide hydrolase and alfatoxin B1 inhibitor, activities also reported to be triggered by bromobenze treatment in rats, already at medium bromobenzene doses, differently from most differently expressed genes which responded only to the highest dose.

To understand the added value of ASCA-genes above other MSTC analysis methods we compared our approach with four different methodologies described for the analysis of time course microarray data: two clustering methods, SOTA and *K*-means, widely used in many gene expression profiling studies, and two hypothesis testing based approaches, timecourse and maSigPro, especially conceived for time series data. For this, the toxicogenomics dataset was analyzed with each of the alternative methods and results were compared to ASCA-genes in terms of provided overall information and biological meaning of feature selection.

The first noticeable difference when comparing ASCA-genes to the methods described earlier, is the absence in the latter of an explicit view of major gene expression features related to the experimental factors. While ASCA-genes gives at first glance a representation of the main gene expression changes occurring along time and across series (Figs 6 and 7), the extraction of this information in the other methods is at least not immediate. The SOTA analysis provided a tree of 26 end nodes in which the gene expression profiles on the different time series and their relative importance were difficult to reveal (Supplementary Fig. S1). Similar conclusions were drawn from the *K*-means partitioning (Supplementary Fig. S2). On the other hand, timecourse and maSigPro packages produced lists of genes with an associated value of statistical significance for differential expression and again without a summarizing representation of experimental factor effects related to major gene expression trends. Next, for comparing the feature selection aspect of ASCA-genes, a selection of genes has to be made for the alternative approaches. In the case of the clustering methods this implied the selection of 'important' clusters and here we encountered the difficulty of not having an objective tool for this selection. We therefore made a selection on the basis of visual inspection of the cluster profiles, selecting

Table 3. Gene selection comparison between ASCA-genes and alternative methods. GO term enrichment analysis of differential set was done with the software Blast2GO (Conesa *et al.*, 2005)

ASCA-genes (373)	SOTA (155)	K-means (136)	Timecourse (256)	maSigPro (155)
Common with ASCA-genes	106	106	119	132
ASCA-genes versus alternative method	Biosynthetic process Translation Fibrinogen complex RNA binding	Lipid metabolic process Mono-oxygenase activity Oxidoreductase activity Retinoid Metabolic process Steroid dehydrogenase activity Response to xenobiotic stimulus Glutathione transferase activity	Microsome Fibrinogen complex Ribonucleoprotein complex Cytoplasm Biosynthetic process Translation Lipid metabolic process	Oxidoreductase activity Steroid biosynthetic process Fibrinogen complex
Alternative method versus ASCA-genes	None	Ribosome	None	None

Numbers in brackets indicate the selection of genes in each method.

according to a series associated pattern of differential expression and by the magnitude of the change. In the case of SOTA, this resulted in the selection of the upper 11 and lower 9 clusters, summing up a total of 155 genes. Upper clusters corresponded basically to genes induced by high and medium doses of the drug, while the lower nine clusters contained repressed genes (Supplementary Fig. S1). In the case of *K*-means, this resulted in the selection of five clusters (136 genes) which also included up and down regulated genes at high and medium bromobenzene doses (Supplementary Fig. S2). In the case of maSigPro, a selection of 155 genes was obtained when applying a significance level of 0.01. Timecourse, on the contrary, does not provide a selection of significant features but a rank of genes ordered by the value of the *MB*-statistic. Taking genes with a positive value for the *MB*-statistic, a total of 256 genes were selected for comparisons. It should be mentioned that the results of the timecourse approach changed with the scale of the data: when data values were multiplied by 100, the gene associated *MB* values greatly increased, becoming positive in many cases (data not shown), which drastically affects the results of any absolute feature selection criterion. This did not happen with any of the other statistical approaches.

Table 3 shows the results of comparing the ASCA-genes feature selection with the alternative approaches. A complete list of selected genes by the different approaches is given in Supplementary Table S1. In all cases, ASCA-genes provided a greater gene selection. Moreover, 70–85% of the genes selected with the alternative method were present in the ASCA-gene pool, except for the timecourse method where overlap was only of 50%. Analysis of the biological meaning content in the groups of genes that were different in each pairwise comparison indicated the presence of significantly enriched GO terms in any of the groups of genes selected by ASCA-genes and not detected by the other approaches. In contrast, for the opposite comparison no enriched GO terms could be detected, except for the *K*-means versus ASCA-genes comparison which revealed the presence of ribosomal proteins not selected by the ASCA-genes approach.

Taken together, these results suggest a stronger discovery power of the multivariate approach and illustrate another distinct feature of the ASCA-genes, i.e. categorization according to values of leverage and SPE statistics. High leverage and low SPE genes are genes that vary according to the main trend and correspond to major molecular functions affected by the treatment. High SPE genes are model diverging data and would correspond to responsive genes with a minority pattern. Low leverage genes show low variance and encode functions less specific in the bulk response. Although the commonality or not of a given gene expression pattern can also be derived through cluster analysis (i.e. high SPE genes can be found in the two most upper clusters of the SOTA result), the explicitness and magnitude of this kind of information is best obtained in the ASCA approach.

4 DISCUSSION

This article addresses the application of ASCA-genes for the analysis of MSTC microarray data. ASCA combines ANOVA and SCA techniques. Basically, ASCA applies PCA to the estimated parameters in each source of variation of an ANOVA model. The method estimates two gene statistics, leverage and SPE that provide information on the adequateness of the model for each gene. This methodology analyses data from a multivariate prospective, taking into account the experimental design and focusing on the sources of variability associated to each of the experimental factors.

The application of the ASCA approach to transcriptomics data has two practical uses. On the one hand, the analysis of the score profiles of the components of the submodels of interest (time, experimental groups and interactions) helps to understand the shared behaviors of gene expression under the studied conditions. On the other hand, the study of the loadings allows the identification of the genes that follow the discovered

patterns. Finally, by analyzing the residuals small group of genes following different expression profiles as those modeled by the ASCA-model can be identified.

One follow up question is how to identify most interesting genes in such approach. We have proposed a criterion for feature selection based on the combined use of two statistics: the leverage, as a measure of the importance of a gene contribution to the multivariate fitted ASCA-model, and the SPE as an evaluation of the goodness of fit of the model to a particular gene. Threshold values for these parameters can be derived by resampling methods or using a weighted χ^2-distribution proposed by Box (1954). The simulated data shows that we can use these measures to categorize genes as (1) genes with a variability pattern associated with the main effects of the experimental factors, (2) genes that diverge from the most abundant behavior but can display a minor gene expression response and (3) genes which basically do not respond to the factors evaluated in the study. Taking as 'interesting' genes those falling into the first two classes, our simulation example also showed that these selection criteria provide specificity and sensitivity values above 90%. Furthermore, the results obtained with experimental microarray data, show that the score profiles derived by ASCA-genes are consistent in the context of a dose-related toxicological response and that the feature selection provided by the leverage and SPE parameters is biologically significant. Comparison of the ASCA approach to other existing methodologies for the analysis of time course data shows the potentials and uniqueness of the proposed method for extracting major and secondary patterns of gene expression associated to the factors of study and for highlighting meaningful gene pools.

Another important benefit of the ASCA-genes approach is the possibility of isolating non-random or unwanted sources of noise. These sources of noise can originate from different labs, experimenters or labeling conditions which associate to only a subset of the data. These non-modeled sources of variability can pollute data and hamper the identification of the signals of interest. We show that when PCA is applied to the gene expression signals in such cases, PCs can pop up that do not associate to the factors of the study and can obscure the interpretation of the results. In fact, complains on the poor interpretability of PCA principal components in microarray analysis has been reported (Raychadhuri *et al*., 2000). As ASCA-genes focuses only on the variability associated to these experimental factors, it clearly outperforms PCA in terms of interpretability and feature selection.

The ASCA methodology relies on standard ANOVA for extracting the variability in the data associated to the experimental factors. ANOVA, however, might not be the most adequate strategy for the analysis of time series as it does not takes into account possible autocorrelation of serial data. In this work we have considered datasets where this problem was not present, as biological and simulated samples were truly independent at each time point. Application of ASCA-genes to longitudinal data, therefore, should be cautious. In any case, the strategy ASCA—PCA restricted to the variability associated to experimental factors—might be extended to other models such as repeated-measures ANOVA or linear mixed models, which would deserve further research.

ACKNOWLEDGEMENTS

This work was funded by the Spanish MCyT (AGL 2003 - 8502-C04-01), INIA (RTA04-013 and RTA05-247) and the Ramón y Cajal Program.

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. *PNAS*, **100**, 10146–10151.

Box,G.E.P. (1954) Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one-way classification. *Ann. Math. Stat*., **25**, 290–302.

Chu,S. *et al*. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.

Conesa,A. *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Conesa,A. *et al*. (2006) maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, **22**, 1096–1102.

Dai,J.J. *et al*. (2006) Dimension reduction for classification with gene expression microarray data. *Stat. Appl. Genet. Mol. Biol*., **5**, Article 6.

Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Appl. Stat*., **28**, 100–108.

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.

Hilsenbeck,S.G. *et al*. (1999) Statistical analysis of array expresion data as applied to the problem of tamoxifen resistance. *J. Natl Cancer I*., **91**, 453–459.

Jansen,J.J. *et al*. (2005) ASCA: analysis of multivariate data obtained from an experimental design. *J. Chemometr*., **19**, 469–481.

Landgrebe,J. *et al*. (2002) Permutation-validated principal components analysis of microarray data. *Genome Biol*., **3**, 19.1–19.11.

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.

Martens,H. and Næs,T. (1989) *Multivariate Calibration*. John Wiley & Sons, Ltd., Chichester.

Nguyen,D. and Rocke,D. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Quackenbush,J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet*., **2**, 418–427.

Raychaudhuri,S. *et al*. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput*., **5**, 452–463.

Roden,J.C. *et al*. (2006) Mining gene expression data by interpreting principal components. *BMC Bioinformatics*, **7**, 194.

Smilde,A.K. *et al*. (2005) ANOVA-Simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048.

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**, 3273–3297.

Storey,J.D. *et al*. (2005) Significance analysis of time course microarray experiments. *PNAS*, **102**, 12837–12842.

Tai,Y.C. and Speed,T.P. (2004) A multivariate empirical Bayes statistic for replicated microarray time course data. *Technical report* 667. Department of Statistics, University of California, Berkeley.

Takashima,K. *et al*. (2006) Effect of the difference in vehicles on gene expression in the rat liver–analysis of the control data in the Toxicogenomics Project Database. *Life Sci*., **78**, 2787–2796.

Timmerman,M.E. and Kiers,H.A.L. (2003) Four simultaneous component models of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika*, **86**, 105–122.

Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expresión data. *Bioinformatics*, **17**, 763–774.