BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 15 2012, pages 2004–2007
doi:10.1093/bioinformatics/bts322

Gene expression

Advance Access publication June 4, 2012

Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach

Moysés Nascimento[1,*], Thelma Sáfadi[2], Fabyano Fonseca e Silva[1] and
Ana Carolina C. Nascimento[1]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, Minas Gerais 36570-000 and [2]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, Minas Gerais 37200-000, Brasil

Associate Editor: Trey Ideker

ABSTRACT

Motivation: In a microarray time series analysis, due to the large number of genes evaluated, the first step toward understanding the complex time network is the clustering of genes that share similar expression patterns over time. Up until now, the proposed methods do not point simultaneously to the temporal autocorrelation of the gene expression and the model-based clustering. We present a Bayesian method that considers jointly the fit of autoregressive panel data models and hierarchical gene clustering.

Results: The proposed methodology was able to cluster genes that share similar expression over time, which was determined jointly by the estimates of autoregression parameters, by the average level of expression) and by the quality of the fitted model.

Availability and implementation: The R codes for implementation of the proposed clustering method and for simulation study, as well as the real and simulated datasets, are freely accessible on the Web http://www.det.ufv.br/∼moyses/links.php.

Contact: moysesnascim@ufv.br

Received on November 24, 2011; revised on March 28, 2012; accepted on May 29, 2012

1 INTRODUCTION

Microarray time series (MTS) analysis allows the researcher to characterize set of genes through their longitudinal pattern of expression. According to Schiliep *et al.* (2003), the MTS data analysis methodologies can be divided into two classes. The first one assumes the observations on the expression at each time as independent variables, and so the usual methods such as hierarchical process (Eisen *et al.*, 1998) and the optimization (Tavazoie *et al.*, 1999) can be directly used to cluster genes with similar expression pattern. The second performs clustering based on the set of parameters estimates from specific models, therefore, it is considered more interesting from the statistical and biological viewpoint, since the temporal expression behavior can be taken into account in the clustering.

Among the second class of methods, the dynamic (Ramoni *et al.*, 2002), the hidden Markov (Schiliep *et al.*, 2003) and the B-splines (Bar-Joseph *et al.*, 2003) models deserve special attention. Although these methods are useful, they are not suitable for relatively small experiments, with less than 10 temporal observations per gene (Bar-Joseph, 2004). However, according to Ernst *et al.* (2005), MTS studies are generally characterized by a large number of genes evaluated but with a small number of the temporal expression measures per gene.

In the field of time series, mainly in econometrics, the Bayesian autoregressive panel data model is recommended for situations with large number of small series (Liu and Tiao, 1980), since it provides an increase in accuracy relatively to autoregressive model fit to each one small series.

Considering these advantages of model-based clustering methods and the difficulty of applying them to a large number of small series, here we propose a Bayesian method that considers simultaneously an autoregressive panel data model fit and a hierarchical clustering of the parameter estimates from this model. In addition, we present the detailed R (R Development Core Team, 2011) codes with comments about the proposed method and its application to a MTS dataset extracted from *Saccharomyces cerevisiae* Stanford MicroArray Database.

2 METHODS

The autoregressive panel data model of order p, AR (p), according to Liu and Tiao (1980), is given by as follows:

$$Y_{it} = \mu_i + \sum_{j=1}^{p} \phi_{i(j)} Y_{i(t-j)} + e_{it},\ i=1,2,\ldots,m;\ \ j=1,2,\ldots,p;\ \ t=1,2,\ldots,n, \tag{1}$$

where Y_{it} is the current value of the series i with mean μ_i; $\phi_{i1}, \phi_{i2}, \ldots, \phi_{ip}$ are the autoregression parameters and e_{it} is the residual term, assumed $e_{it} \sim N(0, \sigma_e^2)$.

The approximated likelihood function in matrix notation is given by as follows:

$$L\left(Y|\Phi, \sigma_e^2\right) \propto \sigma_e^{2^{\frac{-m(n-p)}{2}}} \exp\left\{-\frac{1}{2\sigma_e^2}\left(Y - X\Phi\right)^T\left(Y - X\Phi\right)\right\}, \tag{2}$$

where

$$Y = \left[y_{1(p+1)}, y_{1(p+2)}, \ldots, y_{1(n)}, y_{2(p+1)}, \ldots, y_{2(n)}, y_{m(p+1)}, \ldots, y_{m(n)}\right]^T,$$

$$\Phi = \left[\mu_1, \phi_{11}, \phi_{12}, \ldots, \phi_{1p}, \mu_2, \phi_{21}, \ldots, \phi_{2p}, \ldots, \mu_m, \phi_{m1}, \ldots, \phi_{mp}\right]^T \in R^{m(p+1)},$$

[*]To whom correspondence should be addressed.

$$X=\begin{bmatrix} X_1 & 0 & 0 & 0 \\ 0 & X_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & X_m \end{bmatrix} \text{ and } X_i=\begin{bmatrix} 1 & y_{i(p)} & \cdots & y_{i(1)} \\ 1 & y_{i(p+1)} & \cdots & y_{i(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{i(n-1)} & \cdots & y_{i(n-p)} \end{bmatrix},$$

being the dimensions of these two last matrices $m(n-p)\times m(p+1)$ and $(n-p)\times(p+1)$, respectively.

According to the Bayes theorem, the posterior distribution for the parameters of interest$\left(\Phi \text{ and } \sigma_e^2\right)$ is given by as follows:

$$P\left(\Phi,\sigma_e^2|Y\right)\propto L\left(Y|\Phi,\sigma_e^2\right)\times P\left(\Phi|\sigma_e^2\right)\times P\left(\sigma_e^2\right),$$

where $L\left(Y|\Phi,\sigma_e^2\right)$is the likelihood function presented in equation (2), and $P\left(\Phi|\sigma_e^2\right)$ and $P\left(\sigma_e^2\right)$ make a hierarchical Normal—Inverse Gamma prior distribution: $\Phi|\sigma_e^2\sim N_{m(p+1)}\left(\mu,\sigma_e^2 I\right)$ and $\sigma_e^2\sim IG(\alpha,\beta)$.

Under this approach, the Gibbs sampler algorithm was implemented in R software (R Development Core Team, 2011) using the following full conditional posterior distributions:

$$\Phi|\sigma_e^2,Y\sim N_{m(p+1)}\left(\hat{\Phi}_B,\sigma_e^2\Sigma\right),$$

$$\sigma_e^2|\Phi,Y\sim GI\left(\frac{m(n+1)+2\alpha}{2},D+\frac{1}{2}\left(\Phi-\hat{\Phi}_B\right)^T\Sigma^{-1}\left(\Phi-\hat{\Phi}_B\right)\right),$$

where $D=\beta+\left(Y^TY+\mu^TI\mu\right)-\left(X^TY+I\mu\right)^T\left(X^TX+I\right)^{-1}\left(X^TY+I\mu\right)/2$, $\hat{\Phi}_B=\left(X^TX+\ I\right)^{-1}\left(X^TY+I\mu\right)$, $\Sigma=X^TX+I$ and **I** is an identity matrix.

2.1 Iterative method for clustering genes with similar gene expression patterns

The gene clustering was performed using iterative process in which initially a single panel (only one cluster with all genes) from which the parameter estimates from Model 1 were obtained. These estimates were used as input variables in a Ward clustering analysis (Ward, 1963), then for each resulting cluster, the Model 1 was fitted again. This procedure results in a new set of parameter estimates in each Gibbs sampler iteration and consequently in a new clustering output. The number of clusters in each Ward clustering was defined by Mojena's (1977) criterion; therefore, this number can change in each iteration.

Thus, once guaranteed the Gibbs sampler chains convergence, the convergence to the globally optimal solution (the optimal clustering of which the input data should be belonged to) is also guaranteed. Nevertheless, the Gibbs sampler involves drawing random numbers from full conditional posterior distributions, and due to this, the algorithm was run considering three different sets of start values in order to verify if the clustering results were always the same after the convergence.

Figure 1 shows a scheme of the proposed method for genes whose expression series were modeled by an AR(2) panel data model, the simplest multi parametric model. Under this framework, our main goal is that at the end of the algorithm, the resulting clusters contain series with similar gene expression patterns over time, according to the parameter estimates of the given model.

2.2 Implementation

We implemented our program using R statistical computing environment with the computation conducted with a Intel Core 2 Duo E7500 2.0 GHz with 4 GB of RAM.

2.3 Application to simulated data

We evaluated the proposed clustering method performance using 10 simulated datasets, each one with 117 genes divided into five clusters. The genes expression levels were generated over 10 time points using the model described in equation (1). The function *arima.sim* of R software (R Development Core Team, 2011) was used to simulate the gene expression time series for each gene into each cluster as follows:

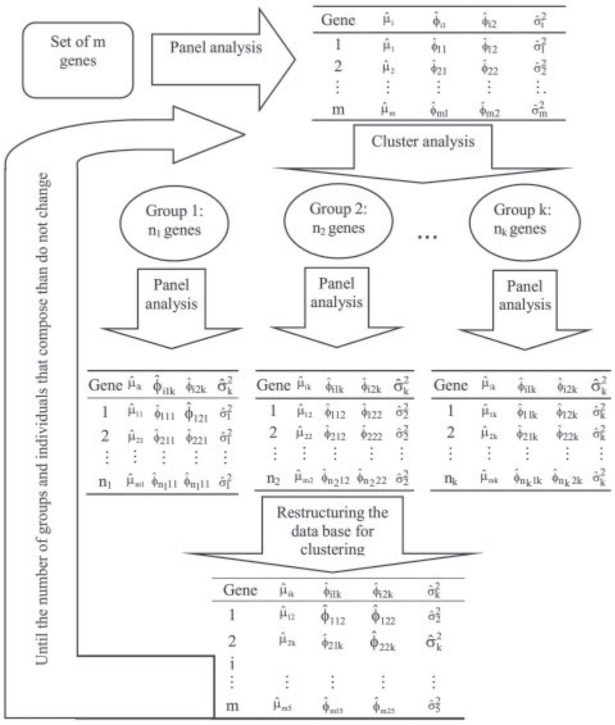
Set of m genes
Panel analysis
Gene $\hat{\mu}_i$ $\hat{\phi}_{i1}$ $\hat{\phi}_{i2}$ $\hat{\sigma}_i^2$
1 $\hat{\mu}_1$ $\hat{\phi}_{11}$ $\hat{\phi}_{12}$ $\hat{\sigma}_1^2$
2 $\hat{\mu}_2$ $\hat{\phi}_{21}$ $\hat{\phi}_{22}$ $\hat{\sigma}_2^2$
m $\hat{\mu}_m$ $\hat{\phi}_{m1}$ $\hat{\phi}_{m2}$ $\hat{\sigma}_m^2$
Cluster analysis
Group 1: n_1 genes
Group 2: n_2 genes
...
Group k: n_k genes
Panel analysis
Panel analysis
Panel analysis
Gene $\hat{\mu}_{ik}$ $\hat{\phi}_{i1k}$ $\hat{\phi}_{i2k}$ $\hat{\sigma}_k^2$
1 $\hat{\mu}_{11}$ $\hat{\phi}_{111}$ $\hat{\phi}_{121}$ $\hat{\sigma}_1^2$
2 $\hat{\mu}_{21}$ $\hat{\phi}_{211}$ $\hat{\phi}_{221}$ $\hat{\sigma}_1^2$
n_1 $\hat{\mu}_{m1}$ $\hat{\phi}_{n_111}$ $\hat{\phi}_{n_111}$ $\hat{\sigma}_1^2$
Gene $\hat{\mu}_{ik}$ $\hat{\phi}_{i1k}$ $\hat{\phi}_{i2k}$ $\hat{\sigma}_k^2$
1 $\hat{\mu}_{12}$ $\hat{\phi}_{112}$ $\hat{\phi}_{122}$ $\hat{\sigma}_2^2$
2 $\hat{\mu}_{22}$ $\hat{\phi}_{212}$ $\hat{\phi}_{222}$ $\hat{\sigma}_2^2$
n_2 $\hat{\mu}_{m2}$ $\hat{\phi}_{n_212}$ $\hat{\phi}_{n_222}$ $\hat{\sigma}_2^2$
Gene $\hat{\mu}_{ik}$ $\hat{\phi}_{i1k}$ $\hat{\phi}_{i2k}$ $\hat{\sigma}_k^2$
1 $\hat{\mu}_{1k}$ $\hat{\phi}_{11k}$ $\hat{\phi}_{12k}$ $\hat{\sigma}_k^2$
2 $\hat{\mu}_{2k}$ $\hat{\phi}_{21k}$ $\hat{\phi}_{22k}$ $\hat{\sigma}_k^2$
n_k $\hat{\mu}_{mk}$ $\hat{\phi}_{n_k1k}$ $\hat{\phi}_{n_k2k}$ $\hat{\sigma}_k^2$
Restructuring the data base for clustering
Gene $\hat{\mu}_{ik}$ $\hat{\phi}_{i1k}$ $\hat{\phi}_{i2k}$ $\hat{\sigma}_k^2$
1 $\hat{\mu}_{12}$ $\hat{\phi}_{112}$ $\hat{\phi}_{122}$ $\hat{\sigma}_2^2$
2 $\hat{\mu}_{2k}$ $\hat{\phi}_{21k}$ $\hat{\phi}_{22k}$ $\hat{\sigma}_k^2$
i
m $\hat{\mu}_{m5}$ $\hat{\phi}_{m15}$ $\hat{\phi}_{m25}$ $\hat{\sigma}_5^2$
Until the number of groups and individuals that compose than do not change

Fig. 1. Flowchart summarizing the Bayesian model-based clustering using autoregressive panel data approach

$$Y_{ik}=\mu_k+\phi_{i1}Y_{i(k-1)}+\phi_{it}Y_{i(k-2)}+e_{ik}$$

where $\boldsymbol{Y}_{ik}$ is the time series (a vector with dimension 1×10) of gene i ($i=1$, 2,..., I_k) of cluster k ($k=1$, 2,..., 5), μ_k is the intercept (average of gene expression), the term *ar* indicates a autoregressive process (Model 1) with parameters ϕ_1 and ϕ_2 (AR(2) process) and σ_e^2 is the residual variance.

The numbers of genes and longitudinal points were chosen taking into account the dimension of the real dataset presented in the next item, and the number of clusters (five) and the values of μ_k, ϕ_{1_k}, ϕ_{2_k} and $\sigma_{e_k}^2$ were determined in according to the results from the real data analysis that will be presented later. For example, the values assumed for ϕ_{1_k} in the simulation were the average values of the estimates obtained from each resulting cluster k, i.e. $\phi_{1_k}=\sum_{\mathrm{i}=1}^{I_k}\hat{\phi}_{\mathrm{i}1_k}\Big/I_k$, where $\hat{\phi}_{\mathrm{i}1_k}$ is the estimate of ϕ_1 for each gene belonging to cluster k and I_k is the number of genes in the cluster k. The values of μ_k, ϕ_{2_k} and σ_{ek}^2 were obtained in same way as ϕ_{1_k}.

In order to compare the clustering method present here with a traditional clustering method, the 10 simulated datasets were also analyzed using k-means algorithm, in which the input variables were the observed gene expression level in each time. Under this approach, given a set of time series gene expression observations: $\boldsymbol{Y}_1=[y_1, y_2,..., y_{10}]$, ... , $\boldsymbol{Y}_{117}=[y_1, y_2, ..., y_{10}]$, the k-means algorithm aims to partition the 117 genes into K ($K\leq 117$) sets $\boldsymbol{S}=[\boldsymbol{S}_1, \boldsymbol{S}_2, ..., \boldsymbol{S}_K]$ so as to minimize the within-cluster sum of squares: $\underset{S}{\arg\min}\sum_{\mathrm{j}=1}^{K}\sum_{x_i\in S_j}\left\|Y_\mathrm{i}-\overline{S}_\mathrm{j}\right\|^2$, where $\overline{S}_j$ is the mean of point in S_j. The choice of k-means is due to its general use in gene expression clustering (Mar *et al.*, 2011; Oh *et al.*, 2011), but it is important emphasize that the success of this method depends of the stated number of clusters, which was the same number used in the simulation study (five), i.e. was considered the optimum condition for the application of this method.

The comparison between Bayesian model-based clustering using autoregressive panel data approach and k-means was evaluated by correct

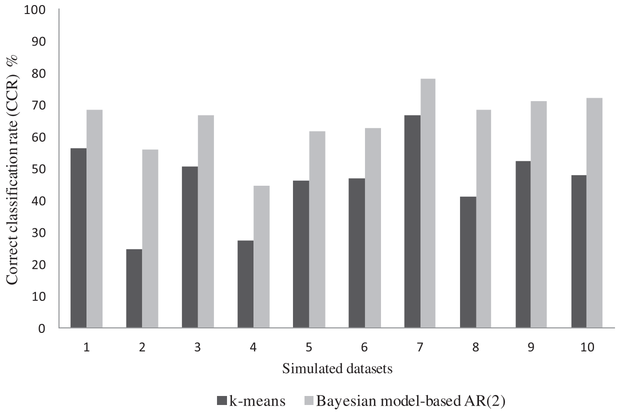

Fig. 2. Correct classification rate for evaluation over 10 replicated datasets

classification rate (CCR), which was computed as the ratio between the numbers of genes clustered into the true cluster (from simulation) and the total number of genes.

2.4 Application to real data

We applied the proposed method to expressions of 117 genes that act on cell cycle of the *Saccharomyces cerevisiae* (Zhu *et al.*, 2000). We used 10 points of the non-synchronized cells data, whose fold-change values came from the expression of mutant strains (treated) compared with wild strains (control), at each evaluated time (0, 15, 30,…, 135 min). At first, the choice by non-synchronized data was a manner to show that the proposed methodology works well for not only for highly selected datasets but also for general class of microarray datasets. For example, if a cell culture is separated out by time, and the expression of genes is measured as a function of time, one can get an idea of which genes vary in expression over time using regression models (Bar-Joseph *et al.*, 2003; Ramoni *et al.*, 2002) with time as a covariate. However, considering that non-synchronized cell populations contain cells at various points in the cell cycle, under this common regression approach, the time effect can be confounded with cell cycle. By the other hand, in the autoregressive regression Model 1, the covariate is not time itself but instead is the expression levels measured at the previous time, thus avoiding the confounding between cell cycle and time. The used dataset may be downloaded at: http://smd.stanford.edu/.

3 RESULTS

3.1 Simulated data

Performance of the clustering methods based on the simulated data is presented in Figure 2, in which the proposed CCR in Section 2.3 was used for evaluation over 10 replicated datasets.

Comparison of the approaches does show that *k*-means clustering performs poorly in relation to Bayesian model-based, once the CCR values were clearly lower. On average, these values were 47.5% (12.56)and 67.42% (9.44), respectively, for *k*-means and Bayesian model-based methods.

Although the simulated dataset was generated under Model 1 assumptions, the comparison between the proposed model-based clustering (using parameter estimates as input variables) and the traditional *k*-means (using observed gene expression level in each time as input variables) methods makes sense. It is because there are biological justifications to dependence between a gene expression at time *t* and previous time, thus whether this dependence is disregarded in the clustering process (as occurred with *k*-means application) really the results will not be satisfactory.

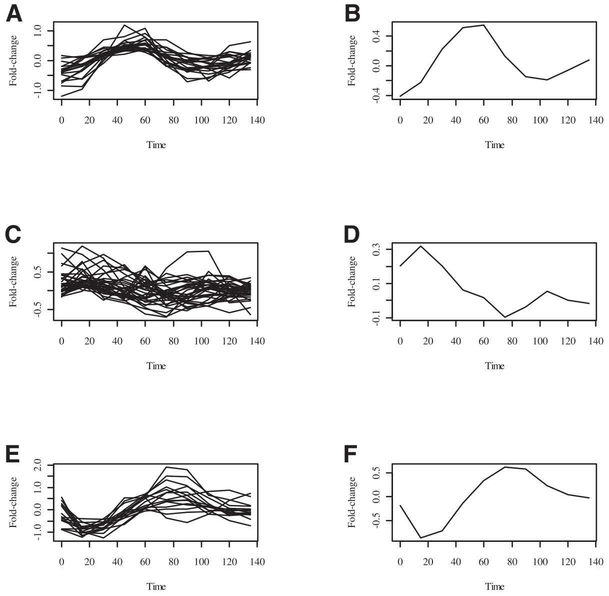

Fig. 3. Time serial expression of three gene clusters found by the proposed methodology. (**A**) Cluster 1 expression profile, (**B**) Cluster 1 average expression, (**C**) Cluster 2 expression profile, (**D**) Cluster 2 average expression, (**E**) Cluster 3 expression profile and (**F**) Cluster 3 average expression

We would like to make clear that the simulation study compared the clustering effectiveness of a new method, which take into account the temporal dependence of gene expression values using Ward algorithm and a traditional method, *k*-means, that considers the independence between these values. Maybe a more honest comparison would be realized between *k*-means and Ward using this same Bayesian model-base approach, but it can be a topic for future research, once the use of *k*-means rather than Ward requires an iterative process (*k*-means optimization) into each Markov Chain Monte Carlo (MCMC) iteration.

3.2 Real data

The algorithm had a running time of $24'46''$ considering Gibbs sampler chains with 10 000 iterations. The chains convergence was accessed by Geweke and Raftery–Lewis diagnostics using *boa* package (Smith, 2007) of R software (R Development Core Team, 2011). These results indicated that the iteration number was enough to ensure the convergence, and furthermore, it was found that the number of clusters (k) and individuals belonging to them stabilized from around 1000 iterations. In addition, it is relevant to highlight that the three different sets of start values provided the same clustering output.

Results allowed genes to be partitioned into five distinct clusters, with 23, 32, 15, 24 and 23 genes in each one. In general, it can be observed that the five clusters have very distinct expression patterns. Among the various differences, it can be noted that the genes that make up Clusters 1 and 2 (Fig. 3) have opposite average behavior during the cell cycle, with observed fold-change values, respectively, for Clusters 1 and 2, negative and positive until a given time, and

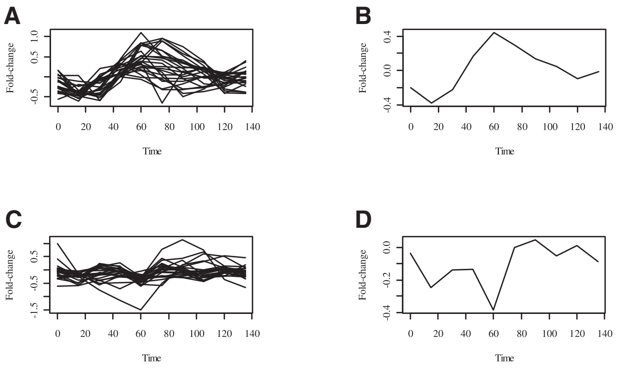

Fig. 4. Time serial expression of two gene clusters found by the proposed methodology. (**A**) Cluster 4 expression profile, (**B**) Cluster 4 average expression, (**C**) Cluster 5 expression profile and (**D**) Cluster 5 average expression

then, after those times the signals are inverted (Fig. 3B and D). Furthermore, genes belonging to Cluster 5 (Fig. 4) generally showed more expression in the control (wild strains) during the cell cycle of *Saccharomyces.*

Figures 3A, C, E and 4A and C depict the gene expression time series in each resulting cluster. The differences in the number of genes that made up each cluster may be related to the number of functions associated with the genes, i.e. clusters with a greater number of genes could be associated with a greater number of functions during the cell cycle.

In summary, we can verify that the proposed methodology was able to cluster genes that share similar expression pattern over time. This similarity can be explained jointly by the magnitude of the dependence between expression at time t and the expressions at time $\tilde{t}1$ and $\tilde{t}2$ (estimates of autoregression parameters, ϕ_1 and ϕ_2), by the average level of expression (μ) and by the quality of the fitted model (σ_e^2), once these values were considered as input variables in the clustering rather than the observed gene expression values over time.

4 CONCLUDING REMARKS

In both simulated and real datasets, the proposed Bayesian model-based clustering method performed very well, once allowed to cluster genes that share similar expression pattern over time. This method was based on the autoregressive panel data model, which is widely used in the econometrics field. Thus, one of the main contributions of this study is to arouse interest of bioinformatician for models from this field in order to describe the time dependence of gene expressions. For example, other powerful models like autoregressive integrated moving average and autoregressive conditional heteroskedasticity can also be considered. Furthermore, high-throughput sequencing data from time series RNA-seq assays can also be analyzed under this approach. However, because to the counting nature of these data (number of reads in a class), other distributions, such as Poisson and Negative Binomial, must be adopted in the likelihood function. With respect to inference methods, the maximum likelihood (Ideker *et al.*, 2000) can be used rather than Bayesian method using generalizations of the Expectation- Maximization (EM) algorithm.

Conflict of Interest: none declared.

REFERENCES

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

Bar-Joseph,Z. *et al.* (2003) Continuous representations of time series gene expression data. *J. Comput. Biol.*, **3**, 341–356.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. Am.*, **95**, 14863–14868.

Ernst,J. *et al.* (2005) Clustering short time series gene expression data. *Bioinformatics*, **21**, 59–168.

Ideker,T. *et al.* (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J. Comput. Biol.*, **7**, 805–817.

Liu,L.M. and Tiao,G.C. (1980) Random coefficient first-order autoregressive model. *J. Econometrics*, **13**, 1980.

Mar,J.C. *et al.* (2011) Defining an informativeness metric for clustering gene expression data. *Bioinformatics*, **27**, 1094–1100.

Mojena,R. (1977) Hierarchical grouping method and stopping rules: an evaluation. *Computer J.*, **20**, 359–363.

Oh,S. *et al.* (2011) Biological impact of missing-value imputation on downstream analyses of gene expression profiles. *Bioinformatics*, **27**, 78–86.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramoni,M.F. *et al.* (2002) Cluster analysis of gene expression dynamics. *Proc. Am. Natl. Acad. Sci.*, **99**, 9121–9126.

Schiliep,A. *et al.* (2003) Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*, **19**, 264–272.

Smith,B.J. (2007) Boa: an R Package for MCMC output convergence assessment and posterior inference. *J. Stat. Software*, **21**, 1–37.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.

Ward,J.H. (1963) Hierarquical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236–244.

Zhu,G. *et al.* (2000) Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, **406**, 90–94.