Genetics and Molecular Research

Bayesian forecasting of temporal gene expression by using an autoregressive panel data approach

M. Nascimento[1]*, F.F. e Silva[2]*, T. Sáfadi[3], A.C.C. Nascimento[1], L.M.A. Barroso[1], L.S. Glória[2] and B. de S. Carvalho[4]

[1]Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[3]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, MG, Brasil
[4]Departamento de Genética Médica, Universidade Estadual de Campina, Campinas, SP, Brasil

*These authors contributed equally to this study.
Corresponding author: M. Nascimento
E-mail: moysesnascim@gmail.com

Genet. Mol. Res. 15 (2): gmr.15027299
Received July 23, 2015
Accepted August 19, 2015
Published June 21, 2016
DOI http://dx.doi.org/10.4238/gmr.15027299

ABSTRACT. We propose and evaluate a novel approach for forecasting gene expression over non-observed times in longitudinal trials under a Bayesian viewpoint. One of the aims is to cluster genes that share similar expression patterns over time and then use this similarity to predict relative expression at time points of interest. Expression values of 106 genes expressed during the cell cycle of *Saccharomyces cerevisiae* were used and genes were partitioned into five distinct clusters of sizes 33, 32, 21, 16, and 4. After removing the last observed time point, the agreements of signals (upregulated or downregulated) considering the predicted expression level were 72.7, 81.3, 76.2, 68.8, and 50.0%,

respectively, for each cluster. The percentage of credibility intervals that contained the true values of gene expression for a future time was ~90%. The methodology performed well, providing a valid forecast of gene expression values by fitting an autoregressive panel data model. This approach is easily implemented with other time-series models and when Poisson and negative binomial probability distributions are assumed for the gene expression data.

Key words: Time series; Temporal gene expression; Posterior predictive distributions

INTRODUCTION

Gene expression time-series analysis allows researchers to characterize a set of genes through their longitudinal pattern of expression (Nascimento et al., 2012). This characterization enables the understanding of different biological processes (functional and regulatory mechanisms) because of the identification of gene groups that share similar expression profiles over time (Mukhopadhyay and Chatterjee, 2007; Korucuoglu et al., 2014).

Given the large number of genes evaluated in microarray and RNA-seq studies, gene clustering is essential to summarize gene expression profiles into a limited number of groups that present similar patterns over time. Among the methods for clustering genes, the dynamic model (Ramoni et al., 2002) and MaTSE (Craig et al., 2013) deserve special attention. However, according to Ernst et al. (2005), such methodologies are not suitable for small experiments where there are few ($N < 10$) temporal observations.

Recently, in order to circumvent this limitation, Nascimento et al. (2012) proposed a joint Bayesian analysis of an autoregressive (AR) model for panel data (Liu and Tiao, 1980) and hierarchical clustering where the autoregression parameter estimates are input variables in the clustering process. Therefore, at the end of the Markov chain Monte Carlo (MCMC) process and after verification of chain convergence, genes belonging to the same group have the same longitudinal behavior.

Although the methodology presented in Nascimento et al. (2012) was not the first Bayesian gene-clustering proposal, according to Dimitrakopoulou et al. (2013), it provided an efficient clustering algorithm by using the flexibility of MCMC methods. Although not yet exploited, this methodology also allows the forecasting of gene expression by means of time-dependence modeling between measures. Thus, the expression level of a gene at a particular time is predicted in an AR process, allowing inferences at non-observed (i.e., future) times. This enables cost reductions associated with gene-expression surveys by using fewer time points and possibly increases the accuracy of analysis by increasing sample size through the inclusion of new observations from this forecasting system. Furthermore, according to Olshen and Jain (2002), this quantitative approach to clustering predicted gene expression is useful for identifying specific genes whose expression is linked with a phenotype (i.e., gene expression changes over time in relation to disease incidence) and for systematic prediction of classes (i.e., clustering genes with similar expression over future times).

The purpose of this work is to propose and evaluate predicted gene expression values by fitting the Bayesian AR panel data model to a microarray time-series dataset extracted from *Saccharomyces cerevisiae*.

MATERIAL AND METHODS

Statistical model

The AR panel data model of order p, AR(p), is given by:

$$\boldsymbol{Y_{it} = \mu_i + \sum_{j=1}^{p} \varphi_{i(j)} Y_{i(t-j)} + e_{it}, i = 1, 2, \ldots, n; j = 1, \ldots, p} \quad \text{(Equation 1)}$$

where Y_{it} is the value of the series *i* with mean, μ_i, $\varphi_{i1}, \varphi_{i2}, \ldots, \varphi_{ip}$ are the autoregression parameters, and e_{it} is the error term ($N(0, \sigma^2)$). The likelihood function, in matrix notation, is given by

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

where

$$Y = \left[y_{1(p+1)}, y_{1(p+2)}, \ldots, y_{1(n)}, y_{2(p+1)}, \ldots, y_{2(n)}, \ldots, y_{m(p+1)}, \ldots, y_{m(n)}\right]^T, \Phi = \left[\mu_1, \varphi_{11}, \varphi_{12}, \ldots, \varphi_{1p}, \mu_1, \varphi_{21}, \ldots, \varphi_{2p}, \ldots, \mu_m, \varphi_{m1}, \ldots, \varphi_{mp}\right]' \in R^{m(p+1)}$$

$$X = \begin{bmatrix} X_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_m \end{bmatrix}_{m(n-p)\times m(p+1)} \quad with\ X_i$$

$$= \begin{bmatrix} 1 & y_{i(p)} & \cdots & y_{i(1)} \\ 1 & y_{i(p+1)} & \cdots & y_{i(2)} \\ & \vdots & \vdots & \ddots \\ 1 & y_{i(n-1)} & \cdots & y_{i(n-p)} \end{bmatrix}_{(n-p)\times(p+1)}$$

The posterior distribution is given by $P(\Phi, \sigma_e^2|Y) = L(Y|\Phi, \sigma_e^2) \times P(\Phi|\sigma_e^2) \times P(\sigma_e^2)$, where $L(Y|\Phi, \sigma_e^2)$ is the likelihood function presented in (2) and $P(\Phi|\sigma_e^2)$ and $P(\sigma_e^2)$ form a hierarchical normal-inverse-gamma prior distribution, $\Phi|\sigma_e^2 N_{m(p+1)}(\mu, \sigma_e^2 I)$ and $\sigma_e^2 \sim IG(\alpha, \beta)$.

The full conditional posterior distributions for Φ and σ_e^2 used to apply the MCMC method (Gibbs sampler algorithm) are given by

$$\Phi, \sigma_e^2|Y \sim N_{m(p+1)}\left(\widehat{\Phi}_B, \sigma_e^2 \Sigma\right) \quad \text{(Equation 3)}$$

$$\sigma_e^2|\Phi, Y \sim IG\left(\frac{m(m+1+2\alpha)}{2}, D + \frac{1}{2}\left(\Phi - \widehat{\Phi}_B\right)^T \Sigma^{-1}\left(\Phi - \widehat{\Phi}_B\right)\right) \quad \text{(Equation 4)}$$

where

$$D = \beta + \frac{(Y^T Y + \mu^T I \mu) - (X^T Y + I\mu)^T (X^T X + I)^{-1} (X^T Y + I\mu)}{2}, \qquad \hat{\Phi}_B = (X^T X + I)^{-1}(X^T Y + I\mu),$$

$\Sigma = X'^T X + I$, and I is an identity matrix.

The R codes for implementation are freely accessible on the web (http://www.det.ufv.br/~moyses/links.php).

Method for clustering genes

Gene clustering was performed through an iterative process that was initially considered one panel with all genes, so the parameter estimates from model 1 were used as input variables using Ward's method (Ward, 1963) for clustering analysis, and for each resulting cluster, model 1 was again fitted (Nascimento et al., 2012). This procedure resulted in new estimates for the parameters, which were once again submitted for cluster analysis. Thus, an iterative process was initiated and repeated until the number of clusters, k, and individuals belonging to them did not show any further changes. The number of clusters was defined using Mojena's criterion (Mojena, 1977). Under this framework, our main goal is that at completion of the algorithm, the resulting clusters contain genes expression patterns over time, according to parameter estimates from the given model. Figure 1 shows a scheme of the proposed method (Nascimento et al., 2012) for genes whose expression series were modeled by an AR(2) panel data model, the simplest multi-parametric model.

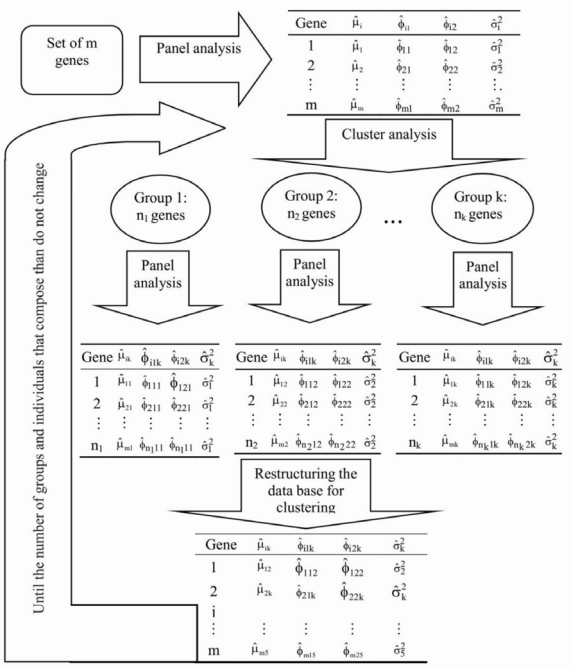

Figure 1. Scheme of the gene clustering method with expression series modeled by a second-order AR panel data model (Nascimento et al., 2012).

Bayesian forecasting in an AR panel data model

After obtaining each cluster by using the iterative process described, the theory of predictive distribution described by Heckman and Leamer (2001) was used to obtain predicted expression values for each gene separately in each cluster (Silva et al., 2011).

Considering the statistical model, AR(p), for panel data from a future value where $e_{i(t+1)} \sim N(0, \sigma_e^2)$, the likelihood function refers to all supposedly independent individuals, *i* (*i* = 1, 2, ..., m), under the matrix form given by

$$L\left(Y_{(t+1)} | \Phi, \sigma_e^2, Y\right) \propto {\sigma_e^2}^{\frac{-m}{2}} exp\left\{-\frac{1}{2\sigma_e^2}\left(Y_{(t+1)} - X\Phi\right)^T\left(Y_{(t+1)} - X\Phi\right)\right\} \quad \text{(Equation 5)}$$

where $Y_{(t+1)} = \left[y_{1(t+1)}, y_{2(t+1)}, \ldots, y_{m(t+1)}\right]^T_{mx1}$, $X = \begin{bmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & X_m \end{bmatrix}_{m \times m(p+1)}$

and $X_i = \left[1, y_{i(t)}, , y_{i(t-1)}, \ldots, y_{i(t+1-p)}\right]_{1 \times (p+1)}$.

Thus, the predictive distribution is given by the following integral:

$$P\left(Y_{(t+1)} | Y\right) \propto \int_{\sigma_e^2}\int_{\Phi} {\sigma_e^2}^{\frac{-m}{2}} exp\left\{-\frac{1}{2\sigma_e^2}\left(Y_{(t+1)} - X\Phi\right)^T\left(Y_{(t+1)} - X\Phi\right)\right\} P(\Phi, \sigma_e^2 | Y)\, d_\Phi d_{\sigma_e^2} \quad \text{(Equation 6)}$$

According to Heckman and Leamer (2001), it is possible to obtain a sample of future observations from the posterior predictive distribution via the MCMC algorithm with the distribution $Y^{(q)}_{(n+1)} | Y \sim N_m(X\Phi^{(q)}, \sigma_e^{2(q)} I)$, where I is an identity matrix of order m(p+1) × m(p+1) and $\Phi^{(q)}$ is a vector of AR parameters at iteration *q*. The point estimate of this value for future observations is given by the mean of this sample, $\hat{P}(Y_{(t+1)} | Y)$.

True value data validation

The data used refer to the expression of 106 genes that act on the cell cycle in *S. cerevisiae* (Zhu et al., 2000) and their Bayesian information criterion values suggest a second-order AR as the most plausible model. This criterion was chosen as the simplest multiparametric model. We used 10 points consisting of fold-change values from the gene expression of mutant strains (treated) as compared to wild-type strains (control), along with each evaluated time point (0, 15, 30, ..., 135 min). The dataset can be downloaded from the *S. cerevisiae* database (http://smd.stanford.edu/).

After determining groups by using the iterative method for gene clustering and applying the predictive posterior distribution approach, we assessed (group-specific) model predictive ability by using mean square error (MSE). To achieve this, we removed the last observation (gene expression at time point 135 min) to allow direct comparison between the predicted values and the observed true values. The formula for calculating MSE for the *j*th group is

$$MSE_j = \frac{1}{ng_j}\sum_{i=1}^{ng_j}\left(Y_{i,135} - \hat{Y}_{i,135}\right)^2$$

where ng_j is the number of observations, $Y_{i;135}$ is the observed value at time point 135 min of the i^{th} series, and $\hat{Y}_{i;135}$ is the predicted value for time point 135 min of the i^{th} series.

Additionally, we calculated the percentage of credibility intervals that contain a gene expression true value, corresponding to time point 135 min ($Y_{i;135}$) and the percentage of agreement between the true values of the last observation (Y_{135}) and their estimates ($\hat{Y}_{135}$).

RESULTS

The results, obtained by Raftery-Lewis and Geweke diagnostics (Raftery and Lewis, 1992; Geweke, 1992) and implemented in the boa package (Smith, 2007), indicated that the iteration number was enough to ensure convergence.

Based on these results, genes were partitioned into five distinct clusters having 33, 32, 21, 16, and 4 genes. Among the differences between clusters, it was noted that genes belonging to clusters 1 and 2, in comparison to those in cluster 4, have opposing expression patterns during the cell cycle (Figure 2A, B, and D). The genes in cluster 3 exhibited consistent behavior (i.e., expression unchanged) during the cell cycle (Figure 2C). Moreover, genes in cluster 5 (Figure 2E) exhibited increased expression levels in the control (wild-type strains) at the beginning and end of the cell cycle.

The number of genes appeared to influence the MSE, given its decrease when the number of expression profiles increased in each group (Table 1). This result was expected, since the Bayesian AR panel data model is recommended for situations involving a large number of small series because of the use of all observations in the estimation of individual parameters for each series (Silva et al., 2011; Nascimento et al., 2011; Nascimento et al., 2012). The percentage of credibility intervals that contained the true values of gene expression for time point 135 min was ~90% (Table 1).

The agreement between the true and estimated values indicates the ability of the model to classify genes as upregulated (genes presenting higher expression levels in the treated group) or downregulated (genes presenting higher expression levels in the control group) efficiently. In summary, this model was capable of accurately predicting future expression levels of genes in both the treated and control groups. In this study, the percentages of concordant signals were 72.7, 81.3, 76.2, 68.8, and 50.0% in the five respective clusters, indicating that the model was able to adequately predict the direction of the forecasted values (Table 1).

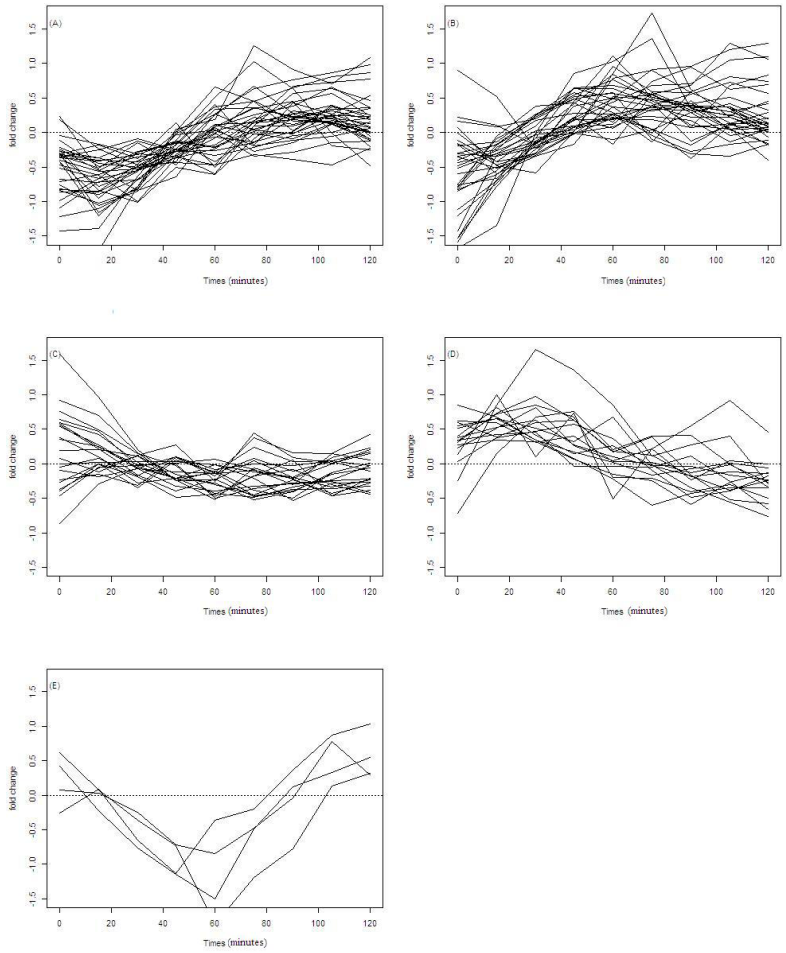

Figure 2. Time-series expression levels of five gene clusters identified using the Nascimento et al. (2012) methodology. **A.** Cluster 1 expression profile. **B.** Cluster 2 expression profile. **C.** Cluster 3 expression profile. **D.** Cluster 4 expression profile. **E.** cluster 5 expression profile. Genes belonging to clusters 1 and 2, in comparison to those in cluster 4, have opposing expression patterns during the cell cycle (A, B, and D). The genes in cluster 3 exhibited consistent behavior (i.e., expression unchanged) during the cell cycle (C). Moreover, genes in cluster 5 (E) exhibited increased expression levels in the control (wild-type strains) at the beginning and end of the cell cycle.

Table 1. Cluster, number of genes (NG), mean square error (MSE), predictive ability (PA), and the agreement of signs (AS).

Cluster	NG	MSE	PA (%)	AS (%)
1	33	0.03	97.0	72.7
2	32	0.08	84.0	81.3
3	21	0.05	95.0	76.2
4	16	0.10	87.5	68.8
5	4	0.20	75.0	50.0
General			87.7	69.8

DISCUSSION

Bayesian forecasting by fitting an AR panel data model is an interesting option, given its high rates of credibility intervals where clusters contain true gene expression values and concordant signals. Therefore, this application of a predictive methodology evaluated over time is a technological innovation that allows the prediction of future gene expression values based on their previous expression patterns.

This modeling method is traditionally used in econometric analysis under situations involving larger numbers of small time series (Liu and Tiao, 1980), represented here, respectively, by the larger number of genes and their expression values over time.

The gene expression values at the final time point were originally present in the observed dataset; however, we removed them in order to evaluate the efficiency of Bayesian forecasting. This result was satisfactory and corresponded to similar results establishing AR model predictive ability. Among these, the work of De Alba (1993) stands out as an extensive revision from time-series methods, where time-series prediction using a fourth-order AR model was simulated and a general efficiency of 75% observed. Hay and Pettitt (2001) obtained 58% in the analysis of twelve pneumonia-incidence time series by using a generalized first-order AR model for counting data. Silva et al. (2011) applied a Bayesian forecasting method by fitting an AR panel data model to longitudinal data relative to the expected progeny difference of beef cattle sires and obtained an efficiency of ~80%.

Finally, we emphasize that this prediction can be generalized to other time-series models, such as AR integrated moving average and AR conditional heteroskedasticity models. Furthermore, this approach is capable of analyzing RNA-seq data by using Poisson and negative binomial distributions (Nascimento et al., 2012).

CONCLUSIONS

The proposed Bayesian framework for predicting gene expression levels worked well in our study, based on a predictive success rate of ~90%.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Foundation for Support of Universidade Federal de Viçosa (FUNARBE), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, #APQ 00825), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERENCES

Craig P, Cannon A, Kukla R and Kennedy J (2013). MaTSE: the gene expression time-series explorer. *BMC Bioinformatics* 14 (Suppl 19): S1. http://dx.doi.org/10.1186/1471-2105-14-S19-S1

De Alba E (1993). Constrained forecasting in autoregressive time series models: a Bayesian analysis. *Int. J. Forecast.* 9:

95-108. http://dx.doi.org/10.1016/0169-2070(93)90057-T
Dimitrakopoulou K, Vrahatis AG, Wilk E, Tsakalidis AK, et al. (2013). OLYMPUS: an automated hybrid clustering method in time series gene expression. Case study: host response after Influenza A (H1N1) infection. *Comput. Methods Programs Biomed.* 111: 650-661. http://dx.doi.org/10.1016/j.cmpb.2013.05.025
Ernst J, Nau GJ and Bar-Joseph Z (2005). Clustering short time series gene expression data. *Bioinformatics* 21 (Suppl 1): i159-i168. http://dx.doi.org/10.1093/bioinformatics/bti1022
Geweke J (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bayesian statistics (Bernando JM, Berger JO, David AP and Smith AFM, eds.). Oxford University, New York, 625-631.
Hay JL and Pettitt AN (2001). Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease. *Biostatistics* 2: 433-444. http://dx.doi.org/10.1093/biostatistics/2.4.433
Heckman J and Leamer E (2001). Handbook of Econometrics. Elsevier Science, Amsterdam.
Korucuoglu M, Isci S, Ozgur A and Otu HH (2014). Bayesian pathway analysis of cancer microarray data. *PLoS One* 9: e102803. http://dx.doi.org/10.1371/journal.pone.0102803
Liu LM and Tiao GC (1980). Random coefficient first-order autoregressive model. *J. Econ.* 13: 305-325. http://dx.doi.org/10.1016/0304-4076(80)90082-2
Mojena R (1977). Hierarchical grouping method and stopping rules: an evaluation. *Comput. J.* 20: 359-363. http://dx.doi.org/10.1093/comjnl/20.4.359
Mukhopadhyay ND and Chatterjee S (2007). Causality and pathway search in microarray time series experiment. *Bioinformatics* 23: 442-449. http://dx.doi.org/10.1093/bioinformatics/btl598
Nascimento M, Sáfadi T and Silva FF (2011). Aplicação da análise de agrupamento para dados de expressão gênica temporal em dados em painel. *Pesquisa Agropecu. Bras.* 46: 1489-1495.
Nascimento M, Sáfadi T, Fonseca e Silva F and Nascimento ACC (2012). Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach. *Bioinformatics* 4: 1-5. http://dx.doi.org/10.1093/bioinformatics/bts322
Olshen AB and Jain AN (2002). Deriving quantitative conclusions from microarray expression data. *Bioinformatics* 18: 961-970. http://dx.doi.org/10.1093/bioinformatics/18.7.961
R Development Core Team (2015). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at [http://www.r-project.org]. Accessed January 2015.
Raftery AL and Lewis SM (1992). Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Stat. Sci.* 7: 493-497. http://dx.doi.org/10.1214/ss/1177011143
Ramoni MF, Sebastiani P and Kohane IS (2002). Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA* 99: 9121-9126. http://dx.doi.org/10.1073/pnas.132656399
Silva FF, Sáfadi T, Muniz JA, Rosa GJM, et al. (2011). Bayesian analysis of autoregressive panel data model: application in genetic evaluation of beef cattle. *Sci. Agric.* 68: 237-245. http://dx.doi.org/10.1590/S0103-90162011000200015
Smith BJ (2007). boa: an R package for MCMC output convergence assessment and posterior inference. *J. Stat. Softw.* 21: 1-37. http://dx.doi.org/10.18637/jss.v021.i11
Ward JH (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58: 236-244. http://dx.doi.org/10.1080/01621459.1963.10500845
Zhu G, Spellman PT, Volpe T, Brown PO, et al. (2000). Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature* 406: 718-722. http://dx.doi.org/10.1038/35021046