BATS: a Bayesian user-friendly software for Analyzing Time Series microarray experiments.

Claudia Angelini,
*Instituto per le Applicazioni del Calcolo, ‘Mauro Picone’, CNR-Napoli, Italy*

Luisa Cutillo
*Telethon Institute of Genetics and Medicine, Napoli, Italy*

Daniela De Canditiis
*Instituto per le Applicazioni del Calcolo, ‘Mauro Picone’, CNR-Roma, Italy*

Margherita Mutarelli,
*Istituto di Scienze dell’Alimentazione, CNR-Avellino, Italy*

Marianna Pensky
*Department of Mathematics, University of Central Florida, USA*

Abstract

Summary:

BATS is a user-friendly software for Bayesian Analysis of Time Series microarray experiments based on the novel, truly functional and fully Bayesian approach proposed in Angelini *et al.* (2006). The software is specifically designed for time series data. It allows an user to automatically identify and rank differentially expressed genes and to estimate their expression profiles. BATS successfully manages various technical difficulties which arise in microarray time-course experiments, such as a small number of observations, non-uniform sampling intervals, and presence of missing or multiple data. BATS can carry out analysis with both simulated and real experimental data. It also handles data from different platforms.
Availability:

BATS is written in Matlab and executable in Windows (Macintosh and Linux version are currently under development). It is freely available upon request from the authors.

1 Introduction

Any biological process (e.g. pharmacological treatment) is a dynamic process which causes a change of gene expression levels in time. Modern microarray technology allows one to measure the expression levels of virtually all genes present in a cell, thus obtaining a "molecular picture" of the cell state. However, the potential of this technique to describe evolution of the cell in time has not yet been fully exploited since only a few methods in statistical genetics take into account the temporal relationship between samples. In fact, most of the existing software packages essentially use techniques designed for static data, so that the results of microarray data analysis are often invariant under permutation of the time points. For example, the SAM software package (see Tusher et al. 2001) was recently adapted to handle time course data by regarding different time points as different groups, while the ANOVA approach by Kerr et al. (2000) was applied to time course experiments by treating the time variable as a particular experimental factor. Papers by Park et al. (2003), Conesa et al. (2006), DiCamillo et al. (2005) and the Limma package by Smyth (2005) have similar approaches. On the other hand, most classical time series or signal processing algorithms have rigid requirements on data (high number of timepoints, uniform sampling intervals, absence of replicated or missing data) which microarray experiments rarely meet.

The past few years saw new developments in the area of analysis of time-course microarray datase (see e.g. de Hoon et al. (2002), Bar-Joseph Z (2004), and most comprehensive approaches of Tai and Speed (2006) and Storey et al. (2005) (implemented by software EDGE by Leek et al. 2006). The novel statistical user-friendly software BATS described in this paper is also specifically designed for analysis of time series microarray data. BATS implements the truly functional fully Bayesian approach of Angelini et al. (2006). It allows a user not only to automatically identify and rank differentially expressed genes, but also to estimate their expression profiles. BATS successfully manages various technical difficulties such as a small number of observations, non-uniform sampling intervals, and presence of missing or multiple data.
2 Methodology

BATS is designed for the data which consists of the records on $N$ genes and describes the difference in genes expression levels between treatment and control. Each record is modeled as a noisy measurement of a function $s_i(t)$ at a time point $t^{(j)} \in [0, T]$:

$$
\hat{z}_i^{j,k} = s_i(t^{(j)}) + \xi_i^{j,k}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, n; \quad k = 1, \ldots, k_i^{(j)}.
$$

Here the number of time point is relatively small ($n \approx 10$), very few replications at each time point are available ($k_i^{(j)} = 0, \ldots, K$, $K = 1, 2$ or 3 ) while the number of genes is very large ($N \approx 10,000$). The objective is to identify the genes which show different functional expressions between treatment and control (i.e. $s_i(t) = 0$), and then to evaluate the effect of the treatment (i.e., estimate $s_i(t) \neq 0$).

For each gene $i$, we expand its expression profile $s_i(t)$ into series over some standard orthonormal basis on $[0, T]$ with coefficients $c_i^{(l)}$, $l = 0, \cdots, L_i$. Legendre polynomials and Fourier basis suitably rescaled and normalized in $[0, T]$ are supported in the current version of BATS.

Following Angelini et al. (2006), genes are treated as conditionally independent and $z_i = D_i c_i + \xi_i$. Here, $D_i$ is the block design matrix, the $j$-row of which is the block vector $[\phi_0(t_j) \phi_1(t_j) \cdots \phi_{L_i}(t_j)]$ replicated $k_j^i$ times; $z_i = (z_{i1}^{1,1} \ldots z_i^{1,k_1} \ldots z_i^{n,1} \ldots z_i^{n,k_n})^T$, $c_i = (c_i^0, \ldots, c_i^{L_i})^T$ and $\xi_i = (\xi_i^{1,1}, \ldots, \xi_i^{1,k_1}, \ldots, \xi_i^{n,1}, \ldots, \xi_i^{n,k_n})^T$ are, respectively, the column vectors of all measurements for gene $i$, the coefficients of $s_i(t)$ in the chosen basis, and random errors. The following hierarchical model is imposed on the data:

$$
\begin{align*}
  z_i & \mid L_i, c_i, \sigma^2 \sim N(D_i c_i, \sigma^2 I_{M_i}) \\
  L_i & \sim \text{Truncated Poisson} (\lambda, L_{\text{max}}) \\
  c_i & \mid L_i, \sigma^2 \sim \pi_0 \delta(0, \ldots, 0) + (1 - \pi_0) N(0, \sigma^2 \tau_i^2 Q_i^{-1})
\end{align*}
$$

All parameters in the model are treated either as random variables or as nuisance parameters which are recovered from the data. Noise variance $\sigma^2$ is assumed to be random, $\sigma^2 \sim \rho(\sigma^2)$ in order to account for possibly non-Gaussian errors which are quite common in microarray experiments. Currently, BATS supports three types of priors: delta-type prior $\rho(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$, the inverse Gamma prior $\rho(\sigma^2) = IG(\gamma, b)$ and the exponential type prior $\rho(\sigma^2) = c_\rho \sigma^{M_i-1} e^{-\sigma^2 \mu_i / 2}$ which lead to normal, Student $T$ and double-exponential errors, respectively. The choice of differentially expressed genes is made on the basis of Bayes factors which are used for multiplicity control and are computed using the novel procedure of Abramovich and Angelini (2006). Once significant genes are detected, the coefficients $c_i^{(l)}$ and, subsequently, the curve $\hat{s}_i(t)$ are estimated by the posterior means. Hyperparameters
\( \pi_0 \) and \( \sigma_0^2 \), \( \gamma \), \( b \) or \( \mu \) are estimated from the data, or can be entered as known. Gene specific parameters \( \tau_i^2 \) and \( L_i \) are estimated by maximizing, respectively, the marginal likelihood and the posterior mean or mode. The advantage of the Bayesian model described above is that all evaluations are carried out in analytic form (see Angelini et al. (2006) for details) which leads to very efficient computations.

3 BATS

BATS is a user-friendly software written in Matlab. Executable program can be obtained from the authors upon request.

BATS supports two main applications: Simulations and Analysis (see Figure 1). Analysis allows an user to apply the methodology described above to either synthetic or real data-set. The data can be analyzed with any of the three error types. Analysis of the human breast cancer data-set by Cicatiello et al. (2004) is provided as guided example. Simulations enable to generate and save synthetic data, and then use it for further analysis. Performance of the technique is evaluated using FDR, FNR, average numbers of correctly detected, missed or missclassified genes and some other standard measures. This feature can be used for choosing experimental design without performing an actual experiment, e.g., one can find an acceptable balance between the cost and the benefits when the increasing number of arrays can be used. In addition, Utilities provide a set of procedures which help to process the input and the output files.

4 Conclusions

This paper describes BATS, a novel statistical user-friendly software specifically designed for analysis of time series data. BATS allows an user to analyze time series microarray experiments having possibly non-Gaussian errors and as few as 10 time points per gene. It is very computationally efficient, since all calculations are based on analytic expressions.
BATS automatically manages non-uniform sampling intervals, and the presence of missing or multiple data. The method accounts for multiplicity, and also selects and ranks differentially expressed genes. To the best of our knowledge, BATS is the only software which estimates and plots gene expression profiles, and provides a “Simulations” utility which allows a biologist to choose among several competitive experimental designs.

BATS can be used to analyze data produced with any microarray platform. The current version of BATS is designed for a one sample problem. However, the two sample applications will be added in subsequent releases.

Acknowledgment

We are grateful to Umberto Amato, Angelo Facchiano and Alessandro Weisz for their constructive comments. This work was supported in part by CNR-Bioinformatics Project, Progetto Oncoproteomica Conv.n.527B/2A/10, and by US National Science Foundation grant DMS-0505133.

References

[Abramovich et al, 2006] Abramovich, F. and Angelini, C. (2006), Bayesian Maximum a Posteriori Multiple Testing Procedure, Sankhya, 68, 436–460.

[Angelini et al, 2006] Angelini, C., De Canditiis, D., Mutarelli, M., Pensky., M., (2006) Bayesian approach to estimation and testing in time course microarray experiments, Tech. Rep. IAC-CNR n. 317/06. Available http://www.na.iac.cnr.it/rapporti/anno2006.htm

[Bar–Joseph, 2004] Bar–Joseph Z., (2004). Analyzing time series gene expression data. Bioinformatics, vol. 20, no 16, 2493–2503.

[Cicatiello et al, 2004] Cicatiello, L., et al. (2004). A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. Journal of Molecular Endocrinology, bf 32, 719–775.

[Conesa et al, 2006] Conesa, A., Nueda, M.J., Ferrer A., and Talon M., (2006). MaSigPro: a method to identify significantly differential expression profiles in time-course microarray-experiments. Bioinformatics, 22, 1096–1102.

[de Hoon et al, 2002] de Hoon, M., Imoto, S., Miyano S., (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines. Bioinformatics, 18, 1477-1485.
[Di Camillo et al, 2005] Di Camillo, B., Sanchez-Cabo, F., Toffolo, G., Nair, S.K., Trajanosky, Z., Cobelli, C., (2005). A quantization method based on threshold optimization for microarray short time series. *BMC Bioinformatics*, 6.

[Kerr et al, 2000] Kerr, M.K., Martin, M. and Churchill G.A., (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7, 819-837.

[Leek et al, 2006] Leek, J.T., Monsen E., Dabney A.R., and Storey J.D., (2006). EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, 22, 507-508.

[Park et al, 2003] Park, T., Yi, S.G., Lee, S., Lee, S.Y., Yoo, D.H., Ahn, J.I. and Lee, Y.S. (2003). Statistical tests for identifying differentially expressed genes in time course microarray experiments. *Bioinformatics*, 19, 694-703.

[Smyth, 2005] Smyth, G.K., (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, 397-420.

[Storey et al, 2005] Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., and Davis, R.W. (2005). Significance analysis of time course microarray experiments (with supplementary material). *Proc. Nat. Acad. Soc.* 12, 12837-12842.

[Tai et al, 2006] Tai, Y.C. and Speed, T.P. (2006) A multivariate empirical Bayes statistic for replicated microarray time course data. *Annals of Statistics* 34, 2387-2412.

[Tusher et al, 2001] V. Tusher, R. Tibshirani and C. Chu (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Soc.* 98, 5116-5121.