

# A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression

Alfredo A. Kalaitzis and Neil D. Lawrence

The Sheffield Institute for Translational Neuroscience, University of Sheffield,  
385A Glossop Road, Sheffield, S10 2HQ, United Kingdom  
{A.Kalaitzis,N.Lawrence}@sheffield.ac.uk

**Abstract.** Two basic forms of analysis recur for gene expression time series: removing inactive (quiet) genes from the study and determining which genes are differentially expressed. Often these analysis stages are applied disregarding the fact that the data is drawn from a time series. In this paper we propose a simple model for accounting for the underlying temporal nature of the data based on a Gaussian process. We present [6] a simple approach for filtering quiet genes, or for the case of time series in the form of expression ratios, quantifying differential expression. We assess via ROC curves the rankings produced by our regression framework and compare them to a recently proposed hierarchical Bayesian model for the analysis of gene expression time-series (BATS). We compare on both simulated and experimental data showing that the proposed approach considerably outperforms the current state of the art.

## 1 Introduction

Gene expression profiles give a snapshot of mRNA concentration levels as encoded by the genes of an organism under given experimental conditions. With the decreasing cost of gene expression microarrays, long time-series experiments have become commonplace, giving a far broader picture of the gene regulation process. Such time series are often irregularly sampled and may involve differing numbers of replicates at each time point. The experimental conditions under which gene expression measurements are taken cannot be perfectly controlled leading the signals of interest to be corrupted by noise, either of biological origin or arising through the measurement process.

As opposed to methods targeted at static experiments (one timepoint), it would seem sensible to consider methods that can account for the special nature of time course data [1, 10, 11]. The analysis of gene expression microarray time-series has benefited the genome-wide identification of direct targets of transcription factors [4, 5] and the full reconstruction of gene regulatory networks [2]. A comprehensive review on the motivations and methods of analysis of time-course gene expression data can be found in [3].

### 1.1 Testing for Differential Expression

A primary stage of analysis is to characterize the activity of each gene in an experiment. Removing inactive or *quiet* genes (genes which show negligible changes in mRNA concentration levels in response to treatments) allows the focus to dwell on genes that have responded to treatment. Removing quiet genes will often have benign effects later in the processing pipeline. However, mistaken removal of profiles can clearly compromise any further downstream analysis. If the temporal nature of the data is ignored, our ability to detect such phenomena can be severely compromised.

This paper, as many other studies, uses data from a *one-sample* setup [1], in which the *control* and *treatment* cases are directly hybridized on a microarray and the measurements are normalized log fold-changes between the two output channels of the microarray, so the goal is to test for a non-zero signal.

A recent significant contribution in regards to the estimation and ranking of differential expression of time-series in a *one-sample* setup is a hierarchical Bayesian model for the analysis of gene expression time-series (BATS) [1], which offers fast computations through exact equations of Bayesian inference, while making a number of prior biological assumptions to accomplish this. In BATS each time-course profile is assumed to be generated from an underlying function, which is expanded on an orthonormal basis (Legendre or Fourier), plus noise. The number of bases and their coefficients, are estimated through analytic computations in a fully Bayesian manner. Thus the global estimand for every gene expression trajectory is the linear combination of some number of bases whose coefficients are estimated by a posterior distribution. In addition, the BATS framework allows for various types of non-Gaussian noise models to be used.

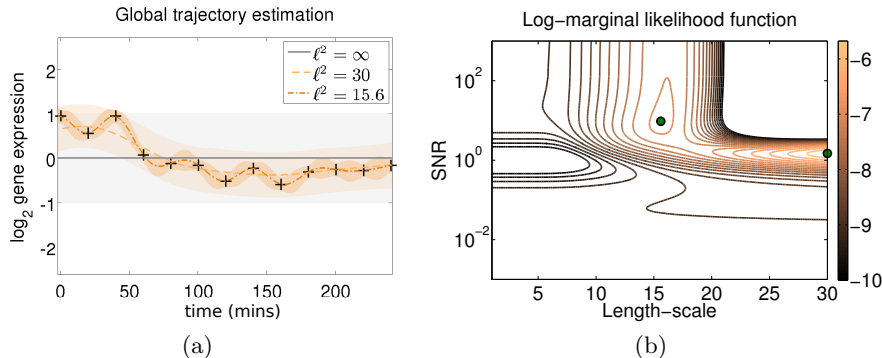
### 1.2 Gene Expression Analysis with Gaussian Processes

*Gaussian processes* (GP) [8] offer an easy to implement approach to quantifying the true signal and noise embedded in a gene expression time-series, and thus allow us to rank the differential expression of the gene profile. In this paper we use the *squared-exponential* covariance function (or RBF kernel). Figure 1 illustrates an example of fitting a GP with an RBF kernel on an experimental profile.

When using different types of models (e.g. with different number of hyper-parameters), a Bayesian-standard way of comparing them is through Bayes factors [1, 9]

$$K = \frac{\int d\boldsymbol{\theta}_1 p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}_1, \mathcal{H}_1) p(\boldsymbol{\theta}_1 | \mathcal{H}_1)}{\int d\boldsymbol{\theta}_2 p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}_2, \mathcal{H}_2) p(\boldsymbol{\theta}_2 | \mathcal{H}_2)},$$

where  $\mathcal{H}_1$  represents the hypothesis where the profile has a significant underlying signal and thus it is truly differentially expressed, and for  $\mathcal{H}_2$  is for no underlying signal in the profile where the observed gene expression is just the effect of random noise.



**Fig. 1.** (a) A GP fitted on the centred profile of the gene *Cyp1b1* (probeID 1416612.at in the *GSE10562* dataset) with different settings of the lengthscale hyperparameter  $\ell^2$ . Crosses are zero-mean hybridised gene expression in time ( $\log_2$  ratios between treatment and control), lines are mean predictions of the GP and shaded areas are the point-wise mean plus/minus two standard deviations (95% confidence region). When the mean function is constant as  $\ell^2 \rightarrow \infty$  (0 inverse lengthscale) then all of the observed variance is attributed to noise ( $\sigma_n^2$ ). When the lengthscale is set to a local-optimum large value ( $\ell^2 = 30$ ), the mean function roughly fits the data-points and the observed variance is equally explained by signal ( $\sigma_f^2$ ) and noise ( $\sigma_n^2$ ). Additionally, the GP has a high uncertainty in its predictive curve. When the lengthscale is set to a local-optimum small value ( $\ell^2 = 15.6$ ) then the mean function tightly fits the data-points with high certainty. The interpretation from the covariance function in this case is that the profile contains a minimal amount of noise and that most of the observed data variance is explained by the underlying signal ( $\sigma_f^2$ ). (b) The contour of the corresponding LML function plotted through an exhaustive search of  $\ell^2$  and signal-to-noise-ratio (SNR) values. The two main local-optima are indicated by green dots and a third local optimum, that corresponds to the constant zero function, has a virtually flat vicinity in the contour, which encompasses the whole lengthscale axis for very small values of SNR (i.e. the lengthscale is trivial if  $\text{SNR} \approx 0$ ).

Depending on the model  $\mathcal{H}$ , these integrals may be intractable. In this paper we present a simple approach to ranking the differential expression of a profile. Instead, we approximate the Bayes factor with a log-ratio of marginal likelihoods

$$K \approx \log \left( \frac{p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}_2)}{p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}_1)} \right),$$

with each likelihood being a function of different configurations of  $\boldsymbol{\theta}$  — the hyperparameters of the RBF kernel. We still maintain hypotheses  $\mathcal{H}_1$  and  $\mathcal{H}_2$  that represent the same notions explained above, but in this case they differ simply by configurations of  $\boldsymbol{\theta}$ . Specifically, on  $\mathcal{H}_2$  the hyperparameters are *fixed* to  $\boldsymbol{\theta}_2 = (\infty, 0, \text{var}(\mathbf{y}))^\top$  to encode a function constant in time ( $\ell^2 \rightarrow \infty$ ), with no underlying signal ( $\sigma_f^2 = 0$ ), which generates a time-series with a variance that can be solely explained by noise ( $\sigma_n^2 = \text{var}(\mathbf{y})$ ). Similarly, on  $\mathcal{H}_1$  the hyperparameters  $\boldsymbol{\theta}_1$  are *initialised* to encode a function that fluctuates in accordance

to a typical significant profile (e.g.  $\ell^2 = 20$ ), with a distinct signal variance that solely explains the observed time-series variance ( $\sigma_f^2 = \text{var}(\mathbf{y})$ ) and with no noise ( $\sigma_n^2 = 0$ ). The log-marginal is then *optimised*, through *scaled conjugate gradients*, with respect to the hyperparameters. The ranking score of a profile is based on how likely  $\mathcal{H}_1$  is in comparison to  $\mathcal{H}_2$ . This methodology is applied on every expression profile in our datasets.

A Gaussian process with an RBF kernel makes the reasonable assumption that the underlying signal in a profile is a *smooth* (infinitely differentiable) function. This property endows the GP with a large degree of flexibility in capturing the underlying signals without imposing strong modeling assumptions (e.g. number of basis functions in BATS) but may also allow it to erroneously pick up spurious patterns (false positives). For a GP approach on *two-sample* data (separate time-course profiles for each treatment), see the work in [9]. GP priors have also been used for modeling transcriptional regulation [7].

## 2 Results and Conclusions

We assume that each gene expression profile can be categorized as either quiet or differentially expressed. As a noisy ground truth, we use data from [4]. For that study, the TSNI algorithm (time-series network identification) was developed to infer the direct targets of TRP63. Furthermore, the direct targets inferred were biologically confirmed by correlation with ChIP-Seq binding regions.

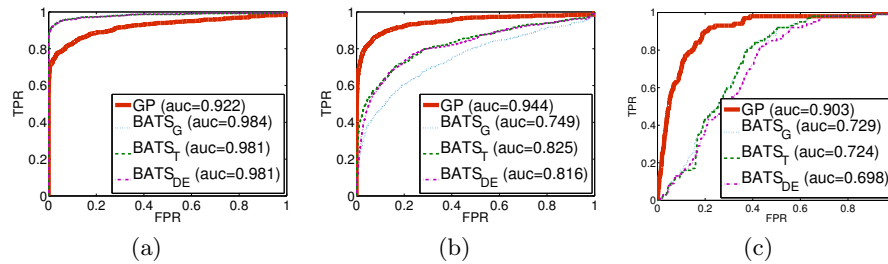
We apply standard GP regression and BATS on two in-silico datasets simulated by BATS and GPs (see Figures 2(a)(b)) and on the experimental data<sup>1</sup>, where only the top 100 ranks of TSNI were labelled as *truly* differentially expressed in the ground truth (see Figure 2(c)). From the output of each model a ranking of differential expression is produced and assessed with ROC curves to quantify how well in accordance to the ground truth (BATS-sampled, GP-sampled, experimental) the method performs.

The experimental data are much more complex and apparently the robust-noise BATS variants now offer no increase in performance. Since the ground truth focuses on the 100 most differentially expressed genes (according to TSNI) with respect to the induction of the TRP63 transcription factor, these results indicate that the proposed approach of ranking indeed highlights differentially expressed genes and that it naturally displays an attractive degree of robustness (similar AUC) against different kinds of noise.

## References

- [1] Angelini, C., De Canditiis, D., Mutarelli, M., Pensky, M.: A Bayesian approach to estimation and testing in time-course microarray experiments. *Stat Appl Genet Mol Biol* 6, 24 (2007)

<sup>1</sup> Available on the GEO database, under accession number GSE10562. Ranking list of direct targets is available for download: [genome.cshlp.org/content/suppl/2008/05/05/gr.073601.107.DC1/DellaGatta\\_SupTable1.xls](http://genome.cshlp.org/content/suppl/2008/05/05/gr.073601.107.DC1/DellaGatta_SupTable1.xls)



**Fig. 2.** One ROC curve for the GP method and three for BATS, using a different noise model (subscript “G” for Gaussian, “T” for Student’s- $t$  and “DE” for double exponential marginal distributions of error), followed by the area under the corresponding curve (AUC). **(a)** Data simulated by BATS, induced with Gaussian noise. Very similar results were acquired for simulated data induced with Student’s- $t$  with 5 degrees of freedom and 3 degrees of freedom. **(b)** On data simulated by GPs. **(c)** On experimental data from [4].

- [2] Bansal, M., Gatta, G.D., Di Bernardo, D.: Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* 22(7), 815 (2006)
- [3] Bar-Joseph, Z.: Analyzing time series gene expression data. *Bioinformatics* 20(16), 2493 (2004)
- [4] Della Gatta, G., Bansal, M., Ambesi-Impiombato, A., Antonini, D., Missero, C., di Bernardo, D.: Direct targets of the TRP63 transcription factor revealed by a combination of gene expression profiling and reverse engineering. *Genome research* 18(6), 939 (2008)
- [5] Honkela, A., Girardot, C., Gustafson, E.H., Liu, Y.H., Furlong, E.E.M., Lawrence, N.D., Rattray, M.: Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences* 107(17), 7793 (2010)
- [6] Kalaitzis, A.A., Lawrence, N.D.: A simple approach to ranking differentially expressed gene expression time courses through gaussian process regression. *BMC Bioinformatics* 12(180) (2011)
- [7] Lawrence, N.D., Sanguinetti, G., Rattray, M.: Modelling transcriptional regulation using Gaussian processes. *Advances in Neural Information Processing Systems* 19, 785 (2007)
- [8] Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA (2006)
- [9] Stegle, O., Denby, K.J., Cooke, E.J., Wild, D.L., Ghahramani, Z., Borgwardt, K.M.: A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology* 17(3), 355–367 (2010)
- [10] Storey, J.D., Xiao, W., Leek, J.T., Tompkins, R.G., Davis, R.W.: Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America* 102(36), 12837 (2005)
- [11] Tai, Y.C., Speed, T.P.: A multivariate empirical Bayes statistic for replicated microarray time course data. *The Annals of Statistics* 34(5), 2387–2412 (2006)