BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 9 2013, pages 1158–1165
doi:10.1093/bioinformatics/btt111

Systems biology

Advance Access publication March 11, 2013

A temporal switch model for estimating transcriptional activity in gene expression

Dafyd J. Jenkins[1,†], Bärbel Finkenstädt[2,*,†] and David A. Rand[1]

[1]Warwick Systems Biology Centre and [2]Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The analysis and mechanistic modelling of time series gene expression data provided by techniques such as microarrays, NanoString, reverse transcription–polymerase chain reaction and advanced sequencing are invaluable for developing an understanding of the variation in key biological processes. We address this by proposing the estimation of a flexible dynamic model, which decouples temporal synthesis and degradation of mRNA and, hence, allows for transcriptional activity to switch between different states.

Results: The model is flexible enough to capture a variety of observed transcriptional dynamics, including oscillatory behaviour, in a way that is compatible with the demands imposed by the quality, time-resolution and quantity of the data. We show that the timing and number of switch events in transcriptional activity can be estimated alongside individual gene mRNA stability with the help of a Bayesian reversible jump Markov chain Monte Carlo algorithm. To demonstrate the methodology, we focus on modelling the wild-type behaviour of a selection of 200 circadian genes of the model plant *Arabidopsis thaliana*. The results support the idea that using a mechanistic model to identify transcriptional switch points is likely to strongly contribute to efforts in elucidating and understanding key biological processes, such as transcription and degradation.

Contact: B.F.Finkenstadt@Warwick.ac.uk

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

Received on November 9, 2012; revised on February 15, 2013; accepted on February 28, 2013

1 INTRODUCTION

One of the archetypal challenges of systems biology is the task of uncovering the network of interactions between genes and proteins using data such as that coming from high-throughput genome-wide technologies or multi-parameter imaging. Time series gene expression data from techniques such as NanoString, reverse transcription–polymerase chain reaction, microarrays or advanced sequencing are particularly valuable for addressing such tasks especially if the system can be perturbed in an informative way. Such data can also be used to get genome-wide understanding of the variation in key biological processes, such as transcription and degradation. In many cases, one is concerned with better-understood systems, such as the circadian clock or cell cycle, where relatively sophisticated models exist. In these cases, it is of interest to uncover both new connections and deeper details of the regulatory interactions. However, when studying systems where there is a much lower density of understanding, one is relatively satisfied with gaining information on the likelihood of the existence of a regulatory interaction or the importance of a regulatory mechanism. Almost all examples studying the response dynamics when systems are subjected to perturbations, such as drug dosing (Eisen *et al.*, 1998) or stress (Windram *et al.*, 2012), or where the progression of disease is studied (Calvano *et al.*, 2005) fall into this latter category.

Analysis of genome-wide time series gene expression data typically involves a number of tasks to parse the time series into groups using various criteria, identify differential expression, select smaller sets of genes for comparative analysis, identify molecular signatures and common regulatory elements, sort the data to identify processes active at certain times and apply network reconstruction algorithms to identify regulatory interactions. One is, therefore, interested in computational approaches to check the similarity or difference in time series expression between genes and conditions. Many techniques for analysing expression profiles have been used [see Androulakis *et al.* (2007) and Bar-Joseph (2004) for overviews], such as hidden Markov models (Schliep *et al.*, 2004; Yoneya and Mamitsuka, 2007), spline functions (Bar-Joseph *et al.*, 2003; Grün *et al.*, 2012) and clustering (Heard *et al.*, 2005; Kiddle *et al.*, 2010). Unfortunately, however, it is relatively rare that, in terms of specific molecular mechanisms, there is much common regulation found across the clusters produced by such methods. This is perhaps less surprising when one notes that the temporal profile of gene expression depends on several processes, such as transcription, degradation and splicing, and that similar profiles can be produced from different combinations of these processes. In particular, the amount of mRNA for a particular gene is the balance between its synthesis and degradation at any point in time. It would, therefore, be helpful if one could identify the effect of these different processes from the data.

This requires the development of algorithms to provide more mechanistic insight by combining time-course expression data with parametric models of gene expression, and there has been some progress in this direction. Relatively sophisticated methods often using stochastic simulation have been developed for extracting parameter estimates from high-resolution time series data (Golightly and Wilkinson, 2011; Komorowski *et al.*, 2009; Toni *et al.*, 2009). However, these approaches are geared towards modelling the intrinsic noise associated with the birth and death

*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.

processes of molecules in single cells and are less suitable for aggregate mRNA data arising from microarray and sequencing experiments. Recently, a flexible parametric model for the response of gene expression to environmental perturbations has been introduced in Chechik and Koller (2009) and can be used in this context. Applying it to gene expression time courses in *Saccharomyces cerevisiae* after diverse environmental perturbations they show that their model, which is based on the product of two sigmoid functions and thus can capture exactly two transitions in the response dynamics, constitutes an improvement over other general functional forms, i.e. polynomials. However, although the magnitude and the timing of the response arise as meaningful parameters, their model is not directly connected to mechanism and is still not general enough to explain a wider range of possible dynamic pattern observed in gene expression, including oscillations.

Consequently, there is a need not only to decouple transcriptional from degradation processes but also to model general forms of transcription in a way that is compatible with the demands imposed by the quality, time-resolution and quantity of the data. In particular, the ability to handle an arbitrary number of transitions where the transcription rate is changed and to infer the number and types of these transitions from such data would clearly be an extremely desirable feature. In this article, we propose an ordinary differential equation model (ODE) model that addresses these issues and at the same time can be effectively fitted to data with sufficient computational efficiency to enable one to handle many genes. It is based on a simple dynamical model of mRNA synthesis and degradation, where transcriptional activity can 'switch' between an arbitrary number of states. The timing and number of transitions, or 'switches', can be estimated efficiently alongside mRNA stability with the help of a reversible jump Markov chain Monte Carlo (RJMCMC) estimation algorithm (Green, 1995). Multiple change-point or switching models have previously been applied to biological systems, such as inferring transcription factor interactions (Opper and Sanguinetti, 2010; Sanguinetti *et al*., 2009), modelling negative feedback in circadian clocks (Aase and Ruoff, 2008) and reconstructing unobserved gene expression dynamics (Finkenstädt *et al*., 2008; Harper *et al*., 2011). However, these models have so far only supported binary expression dynamics, which are not general enough to capture expression dynamics with multiple steady-state expression levels.

The structure of the article is as follows. We first introduce the modelling approach and estimation algorithm. The performance of the algorithm has been studied extensively for artificial data (Supplementary Section 'Simulation study'). To demonstrate the methodology and its potential further uses, we focus on modelling the wild-type behaviour of a set of 200 chosen oscillatory expressed genes of the model plant *Arabidopsis thaliana*. The approach allows us to investigate whether genes with similar switch event times also have correlated promoter motifs. Furthermore, we introduce a Bayesian hierarchical approach to pool data from several experiments and present results for estimation of mRNA stability. The example datasets consist of time series from three experiments (called E1, E2 and E3) of varying timescales and sampling regimes under some mock treatment conditions (see Supplementary Fig. S1 for examples). Each experiment originally consists of >30 000 probes [Sclep *et al*. (2007); www.catma.org], which map >25 000 genes from the TAIR9 genome annotation [Lamesch *et al*. (2012); www.arabidopsis.org]. Here, we focus on a subset of 200 oscillatory genes (chosen according to their correlation to a sine function for the expression data from E1). The set includes a number of 'core' circadian clock genes (such as LHY, AT1G01060; CCA1, AT2G46830; TOC1, AT5G61380). A list of the 200 genes can be found in Supplementary Table S1.

2 A MULTI-SWITCH MODEL AND ITS INFERENCE

We assume that the aggregate dynamics of mRNA over a population of cells can generally be described by a piecewise linear ODE model where because of transcriptional regulatory processes, the transcriptional rate of a gene changes ('switches') from τ_{i-1} to another rate τ_i at time point s_i

$$\frac{dM}{dt} = \begin{cases} \tau_0 - \delta M(t) & \text{for} \quad 0 < t \leq s_1, \\ \vdots & \vdots \\ \tau_k - \delta M(t) & \text{for} \quad s_k < t \leq L. \end{cases} \tag{1}$$

Here, $M(t)$ denotes mRNA concentration at time t, δ is the rate at which mRNA is degraded and L is the length of the time interval over which gene expression is observed. An increase in the transcription rate τ_i from the previous regime can be interpreted as an 'on-switch', whereas an 'off-switch' is associated with a decrease of transcriptional activity. However, we note that the expression might not be fully turned off, and that there may be more than just two states. We will refer to the model in (1) as *switch model*. Neither the location of the *switch-times* $s_1, \ldots, s_k$ nor the number of *switches* k is known and need to be estimated along with the kinetic parameters of the model. Solving the linear ODE for each linear regime and iteratively inserting the final state of a previous regime as initial condition of the next regime one can derive the following general solution

$$\begin{aligned} M(t) = M(0)e^{-\delta t} &+ \frac{\tau_0}{\delta}(1 - e^{-\delta t}) \\ &+ \frac{\tau_1 - \tau_0}{\delta}(1 - e^{-\delta(t-s_1)})I_{t>s_1} + \ldots \\ &+ \frac{\tau_k - \tau_{k-1}}{\delta}(1 - e^{-\delta(t-s_k)})I_{t>s_k} \end{aligned} \tag{2}$$

for an initial condition $M(0)$, where $I_{t>z} = 1$ if $t > z$ is an indicator function. Note that by setting

$$\alpha_0 = M(0), \quad \alpha_1 = \frac{\tau_0}{\delta}, \quad \alpha_2 = \frac{\tau_1 - \tau_0}{\delta}, \ldots.$$

and

$$X_0 = e^{-\delta t}, \quad X_1 = 1 - e^{-\delta t}, \quad X_2 = 1 - e^{-\delta(t-s_1)}, \ldots$$

Equation (2) is a linear model

$$M(t) = \alpha_0 X_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots. + \alpha_{k+1} X_{k+1} \tag{3}$$

for given degradation rate δ and switch-times $s_1, \ldots, s_k$. The dimension of the model is determined by the number of switches k. The case of no interior switch points corresponds to $M(t) = \alpha_0 X_0 + \alpha_1 X_1$, that is the solution of a single linear ODE from an initial condition $M(0)$, whereas each additional switch-point adds another additive term allowing for

convergence to a new equilibrium. Inference is carried out assuming that the 'true' model is unknown but comes from a class of models $M_0, M_1, ..$ where M_k denotes the model with k switching points. Using the notation of (3), each model M_k is associated with a parameter vector $\theta_k = (\alpha_0, \ldots, \alpha_{k+1}, s_1, \ldots, s_k, \delta, M(0))$, the dimension of which changes with the model. Inference about k and θ_k is based on the target distribution that is the joint posterior $p(k, \theta_k)$. As this is a case of changing model dimension, we shall generate samples from the joint posterior using reversible jump Metropolis Hastings (Green, 1995).

Let $\mathbf{y} = (y_{t_i}^{(r)}; i = 1, \ldots, T; r = 1, \ldots, R)$ denote the gene expression data for a particular gene where R denotes the number of replicate time series, each with T observations for a given experimental setting. Note that the notation that each replicate has T observations is only used for simplicity. It will be obvious how to allow for a different number of observations per replicate. Assuming that the residuals between the ODE solution (2) and the expression data are i.i.d. normal with mean zero and unknown variance σ^2, the log likelihood function is

$$l(\theta_k|\mathbf{y}) = -(TR)\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}\sum_{r=1}^{R}(M_k(t_i) - y_{t_i}^{(r)})^2, \quad (4)$$

where θ_k denotes the vector of all unknown parameters, including the initial condition, and $M_k(t_i)$ is the solution to the differential equations defined by Equation (2) under model M_k.

For given switch-times and degradation rates, we have originally devised a complete Bayesian regression approach to the model in (3) by assigning priors and sample from the conditional posterior (Denison *et al*., 1998) of the regression coefficients α_j. However, analogous to the conclusions of Denison *et al.* (1998) in the context of a spline model, we found that the α_j can be calculated by standard least squares regression, which is computationally substantially faster leading to results that, for our purposes, are not distinguishable from the ones obtained from the full Bayesian regression models. Moreover, inference about the actual values of the α_j is not directly of interest, as the time series data can only be assumed to be proportional to the concentration $M(t)$, and the values of α_j are affected by this scaling.

We use a vague gamma prior for the precision, i.e. $p(\sigma^{-2}) = \gamma(10^3, 10^3)$ and update the chain for σ^{-2} via a Gibbs step as in the usual normal Bayesian regression model. The degradation rate δ is a parameter shared by all models and ordinary MCMC updating schemes can be applied. Here, we use a vague normal prior for $log(\delta)$ and a random walk Metropolis updating scheme on the log scale. The prior model for k can be specified by a Poisson distribution $f(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ conditioned on $k \leq k_{max}$ (Green, 1995). Note that by changing λ, the expected number of switches can be controlled reducing potential model overfitting.

With regard to the switch points, we adapted the reversible jump specifications of the algorithms used in Green (1995) and Denison *et al*. (1998) to our model. We assume that the prior switch-time positions $s_1, \ldots, s_k$ are uniformly distributed on $[0, L]$ and classify three possible moves:

(i) movement of a randomly chosen existing switch-point s_i with probability $\eta_k = 1 - b_k - d_k$;
(ii) addition of a switch with probability $b_k = c\ \min\left(1, \frac{f(k+1)}{f(k)}\right)$; and
(iii) deletion of a switch with probability $d_k = c\ \min\left(1, \frac{f(k-1)}{f(k)}\right)$

for some constant $c \in [0, 1/2]$. For $k = 0$, we set $b_0 = 1$ and $d_0 = 0$. The acceptance probability for the moves follows the general rule (Denison *et al*., 1998; Green, 1995)

$$\alpha = \min(1,\ \text{likelihood ratio} \times \text{prior ratio} \times \text{proposal ratio}). \quad (5)$$

For the position change in (i), we randomly chose a switch-time s_i from the k existing switches, and a candidate value $\tilde{s}_i$ is drawn uniformly on $[s_{i-1} + \phi, s_{i+1} - \phi]$, where ϕ is a fixed minimum time between switch-times. The acceptance probability (5) for this move is

$$\alpha = \min\left(1, \text{likelihood ratio} \times \frac{((s_{i+1} - \phi) - \tilde{s}_i)(\tilde{s}_i - (s_{i-1} + \phi))}{((s_{i+1} - \phi) - s_i)(s_i - (s_{i-1} + \phi))}\right),$$

where likelihood ratio here generally refers to the ratio of likelihood of proposed new values of parameters divided by the current likelihood. For move (ii), addition of a switch, we propose a new switch-time s' uniformly on $f(L, s, \phi)$, the support of L given the constraints imposed by ϕ on switch-times s. The proposed value will lie in some interval $[s_i + \phi, s_{i+1} - \phi]$. The prior ratio is then

$$\frac{f(k+1)}{f(k)}\frac{2(k+1)(2k+3)}{f(L, s, \phi)^2}\frac{(s' - (s_i + \phi))((s_{i+1} - \phi) - s')}{(s_{i+1} - \phi) - (s_i + \phi)},$$

the proposal ratio is $\frac{d_{k+1}f(L,s,\phi)}{b_k(k+1)}$ and the acceptance probability for a suggested switch is computed by inserting these into (5). Finally, for move (iii) a switch is chosen randomly from the set of existing switches, and the acceptance probability for this move has the same form as for move (ii) with all ratios inverted.

Figure 1 shows the fit of the switch model to E1 data for the core clock gene LHY (AT1G01060). The algorithm estimates a mean half-life of 1.3 h for LHY mRNA and identifies a total of six switches, which consist of three periodically recurring switches per day. LHY is a core regulating component of the *A.thaliana* clock, and it has been shown to be induced before dusk and has a peak of expression at dawn (Schaffer *et al*., 1998). Around dawn, other clock components repress the expression of LHY resulting in rhythmic expression (Pokhilko *et al*., 2012). The E1 data have a 16:8 h light–dark cycle. The estimated periodic switches show the initial gene induction several hours before dusk and repression at or shortly after dawn.

This example clearly shows that the model is able to identify asymmetric oscillations resulting from unequal length of on and off times and switches that cause additional modes or 'shoulders' in the cyclic patterns. The traces of the RJMCMC algorithm for the switches and their times are plotted in Figure 1C. We have summarized the posterior results by the marginal distribution of all accepted switch-times by fitting a Gaussian mixture model to a function estimated by a non-parametric kernel density. From this, all local maxima are identified and approximated by fitting a mixture of Gaussians. We found this to work well in simulation studies, but note that other parametric or non-parametric approaches could be applied here. Each mode represents a possible switch, and the location is summarized by the mean and the corresponding two-sigma band (shown in Fig. 1C). The latter could be taken as an indicator of the 'strength' of a switch. It should, however, be noted that by summarizing the marginal

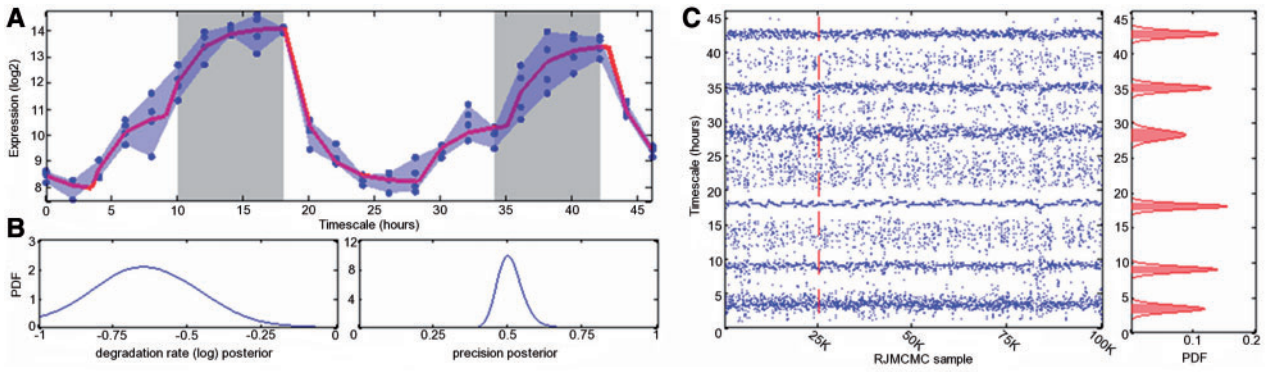

Fig. 1. Output from the RJMCMC for the gene LHY in E1. (**A**) The expression data (individual samples shown by ° with shading between the extreme samples at each time point). E1 data have a 16:8-h light–dark cycle, with the light periods starting at 18 and 42 h and the dark periods starting at 10 and 34 h. Grey-shaded regions indicate periods of no light. The red line shows the fitted piecewise linear ODE model for mean posterior parameter values. Time in hours is on the *x*-axis, and mRNA expression in log-scale is on the *y*-axis. (**B**) The posterior distributions for the sampled degradation rate (left) and precision (shown here in terms of σ) (right). (**C**) The sampled switch-times (left) and the posterior distributions (right) for the six estimated switches, s, and the shaded regions show $s_i \pm 2\sigma_i$. 100 K RJMCMC iterations are generated, and the end of the burn-in period is indicated by the dashed red line at 25 K iterations on (C)

distribution over all accepted switch-times, we are averaging our posterior information over all models entertained by the RJMCMC algorithm. Although we do not follow this in detail, here, we note that it may be useful to investigate the MC traces in more detail for correlation between models.

Convergence for gene expression datasets from E1 is usually achieved after 5 K iterations (Supplementary Fig. S2), and the posterior densities can be estimated from the RJMCMC traces of 75 K iterations, after the first 25 K iterations are discarded as burn-in. The computational time for 100 K iterations was on average 128 s on a 2.8 GHz computer. Fitting this model is thus computationally feasible for thousands of genes and can be easily parallelized.

3 RESULTS

To gain a systematic understanding of how the estimation algorithm performs for time series of varying sampling frequencies and noise levels, we generate synthetic datasets, for chosen kinetic parameter values and using data from E1 to obtain realistic sample sizes and noise levels. The full study benchmarking the performance of the model can be found in Supplementary Section 'Simulation study'. For the synthetic data, we are able to obtain accurate estimates for timing and number of switches and degradation rates under all but the largest noise level (taken as the 95th percentile from the E1 data). We observe that the parameter estimation is invariant of the mode of the switch (increase or decrease transcription rate), that multiple switch points must have at least one observation between them to be reasonably estimated and that a higher sampling frequency is generally more informative for estimation than a larger number of replicate samples. To demonstrate further use of this approach, we now present case studies referring to the 200 example circadian time series.

3.1 Correlation of switch-times with promoter motifs

A common aim of gene expression analysis is to identify potential common regulatory mechanisms between groups of genes through clustering gene expression and enrichment of semantic similarity, such as Gene Ontology (Pesquita *et al.*, 2009), or sequence similarity, such as promoter motif structure (Cooper *et al.*, 2006). Here, we shall use the estimation results from the switch model for the clustering of genes according to similarity in switch-time distributions. Compared with the usual clustering approaches based on similarity of the expression profiles, the basic difference is that we can identify groups of genes that change transcriptional activity around the same times irrespective of mRNA stability and whether such regulation is up or down. Our clustering is based on the similarity matrix whose entries quantify the pairwise distance between the estimated posterior marginal distributions of switch-times (SD) of the genes. Figure 2 shows an example cluster from the 200 circadian genes where we used the symmetric Kullback–Leibler (KL) distance (Kullback and Leibler, 1951). The KL distance is a common choice for computing distances between probability densities, but we note that other distance measures could be implemented in a straightforward way.

Clustering is performed in all cases applying the affinity propagation algorithm (Frey and Dueck, 2007) to the similarity matrix. Multiple similarity matrices can be linearly combined, allowing SDs from multiple experiments to be combined. One can also focus the clustering on subsets of the domain of the SD. If the domain is equal to the total length of observational time, then genes with the same switch-times are clustered together irrespective of whether they are on or off switches. As the genes are all circadian here, the corresponding expression profiles of the genes in that cluster seem to be either in the same or in the opposite 'phase' (Fig. 2A). If the distance measure is applied separately to the on and off times, depending on whether transcription is increased or decreased, then two similarity matrices can be defined for each gene pair. Applying the clustering to the sum of the two resulting similarity matrices gives clusters where the expression patterns are in phase (Fig. 2B). The difference from clustering according to overall similarity of the expression profile is visible in Figure 2C where the red-highlighted example gene is now in a group of more highly correlated expression

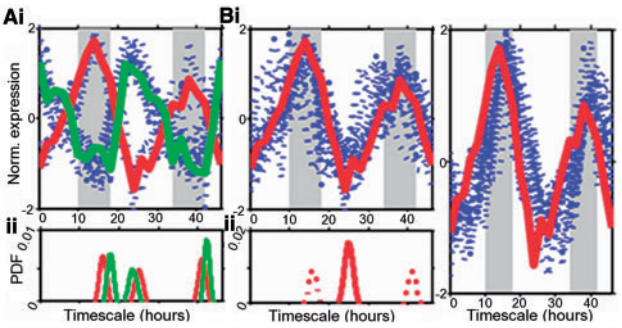

Fig. 2. Example clusters using different similarity scores. (**A**) Cluster containing 21 genes based on the switch-time distribution (SD) over the whole time interval: (**i**) normalized expression profiles, highlighted are two example genes with inverted expression dynamics (red = AT1G10740; green = AT1G12845); (**ii**) SD for the two highlighted genes, indicating the three similar switch times. (**B**) Cluster containing 24 genes based on the sum of the individual similarity matrices obtained by separately considering on and off times of the SD: (**i**) normalized expression profiles, with the same gene (AT1G10740) in red; (**ii**) SD for the highlighted gene where the on set is given by the solid line, and the two off sets are given by the dotted line. (**C**) Cluster containing 22 genes, including the same gene (AT1G10740) in red, resulting from using the sum of squared error between normalized expression profiles as similarity score

profiles, but the timing of its switches is visibly earlier than for most other genes in that cluster.

A commonly explored hypothesis is that correlated gene expression patterns will also have correlated promoter structure and regulation mechanisms. In practice, such correlations have yet to be confirmed. By combining our analysis of switch-time similarity with promoter motif data we ask whether our approach can shed more light on this issue. We investigate how frequently certain listed motifs are encountered in clusters of genes. If a motif has a high frequency for the genes in a given cluster then it is more likely that the corresponding transcription factor binding sites are key for the regulation of those genes. This could give us an indication of which genes may be turned on or off by the same transcription factors. Using position-specific scoring matrix (PSSM) data from the TRANSFAC (Matys *et al.*, 2006) and PLACE (Higo *et al.*,1999) databases, 25 plant motifs were selected, where each motif is present in at least 50 of the 200 circadian gene subset, identified using APPLES (Baxter *et al.*, 2012) (see Supplementary Tables S2 and S3 for motifs). These motifs can be grouped by sequence similarity into three broad classes of promoter motifs, and a similarity matrix is generated from co-occurrence of motifs between each pair of genes, which can then be linearly combined with each of the three similarity scores and clustered.

One could obtain a clustering based on motif co-occurrence alone (Supplementary Fig. S3A), which does not yield any temporal correlation in the resulting expression profiles (Supplementary Fig. S4). On the other hand, clustering only the expression time series (Supplementary Figs S5–S7) brings up some correlated promoter structure when using switch-times, rather than overall expression (Supplementary Fig. S3B). However, the approach is most informative when both the motif co-occurrence and time series information are combined in that we are able to identify strong correlations in promoter structure together with temporal separation of profiles between clusters.

Figure 3 shows heatmaps for the proportions of each motif present in each of the clusters resulting from combining the motif co-occurrence similarity matrix with the similarity score of either the switch-times or the overall expression profiles. The last column (P) gives the proportion of each motif in the overall population of the 200 circadian genes. It can clearly be seen that clusters based on the whole switch-time distribution together with motif co-occurrence show a very high proportion of some motifs in several clusters. For instance, in Cluster 1, all motifs from Group 1 and 2 (19 of 25 motifs) are present in 40% or more of the genes, and in Cluster 7, 11 motifs from Groups 1 and 2 are present in >70% of genes, whereas the population mean motif proportion is only 28% (Fig. 3A). Clusters 1 and 7 contain a number of significantly overrepresented promoter motifs, with $q \leq 0.001$ (denoted by triple asterisk in Fig. 3) using the hypergeometric test and corrected with a false discovery rate of 5% (Benjamini and Hochberg, 1995). Eight clusters contain significantly overrepresented motifs and a clear separation of genes containing motifs from Groups 1, 2 and 3. A similar result is observed using the additively combined separate on and off components and motif co-occurrence for clustering (Fig. 3B), which, naturally, results in a slightly larger number of clusters. Comparing with clustering based on expression profiles together with motif co-occurrence, we find that the proportion of motifs in clusters is generally smaller and fewer significant clusters are observed (Fig. 3C). Plots of the expression profiles separated into clusters are shown in Supplementary Figures S8–S10. Cluster membership and motif proportions for the combined similarity measure clusters are given in Supplementary Tables S4–S7.

This example makes it apparent that clustering based on switch-times combined with motif co-occurrence is useful in identifying correlation between gene expression and promoter structure and, therefore, also in identifying potential regulatory interactions. Motif instance data are often noisy because of redundancy and degeneracy in PSSMs, and in genome-scale expression data sets many genes may share expression profiles, but not regulatory dynamics. By incorporating switch-times with motif data we can link specific temporal events in transcription with specific promoter structures.

3.2 Hierarchical modelling of multiple time series

Provided that the mRNA process exhibits non–steady-state behaviour, the use of the switch model allows for inference on mRNA degradation rates for many genes without having to resort to additional experiments that are specifically targeted at transcriptional inhibition. Although it is straightforward to obtain posterior distributions from one experiment, it may also be of interest to pool the information from several experiments allowing for the possibility that, for a given gene, the mRNA degradation rate across the experiments should be similar but allow for variation because of the setting of the experiment (different laboratories, techniques, mock treatments and time spans). We also wish to incorporate informative prior information from the study by Narsai *et al.* (2007), which gives estimates of mRNA degradation rates of >13 000 *A.thaliana* genes while noting that use of such data is problematic because, by its nature,

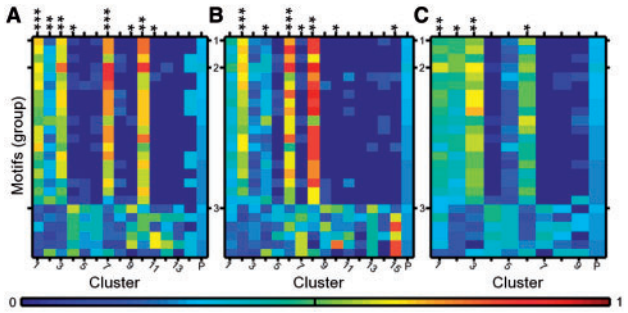

Fig. 3. Heatmaps of proportion of motifs present in gene clusters using different similarity measures. Motifs are aligned on *y*-axis and grouped into three classes by sequence similarity. Each column gives the proportion of motifs in a gene cluster. (**A**) Motif co-occurrence is combined with the switch-time distribution (SD) similarity matrix over the whole time interval, (**B**) motif co-occurrence is combined with the SD similarity separately over on and off times. (**C**) Motif co-occurrence is combined with similarity of the overall expression profiles. Clusters are assigned significance by the hypergeometric test with false discovery rate correction on the cluster motif proportions against the population motif proportions; $* = q \leq 0.05$, $** = q \leq 0.01$ and $*** = q \leq 0.001$. Cluster 'P' shows the population motif proportions

there is no independent validation of the results of such studies, and their experimental conditions are different from ours. We address this in the following Bayesian hierarchical modelling framework.

Let $\mathbf{y}^{(i)}$ denote the time series data (including replicates) for a particular gene in experiment i and $\mathbf{Y}$ the pooled data for that gene over N experiments. Bayesian hierarchical modelling (Gamerman and Lopes, 2006) provides a natural framework to account for the fact that parameters may be similar but are subject to stochastic variability between experiments. Assuming experiments are independent, the full log likelihood is

$$l(\theta; \mathbf{Y}) = \sum_{i=1}^{N} l(\theta_k^{(i)} | \mathbf{y}^{(i)}) \tag{6}$$

where $l(\theta_k^{(i)}|\mathbf{y}^{(i)})$ is the log likelihood as specified in (4), and each $\theta^{(i)}$ is a vector of varying dimension k_i containing all model parameters for experiment i. In a Bayesian hierarchical model, we assume that parameters or subsets of the parameters are similar but subject to some stochastic variation in the sense that they are realizations of the same probability distribution whose parameters we wish to identify. Here, we assume that the mRNA degradation rate for a gene across the experiments comes from the same distribution $\delta^{(i)} \sim p(\delta^{(i)}|\Theta_\delta)$, which is characterized by the parameter vector Θ_δ. The latter thus quantifies the mean value and variability of the degradation rate across the experiments. It should be noted that with respect to the switch-times we maintain a non-hierarchical structure, as regulatory switches may generally occur at different times for each of the experiments given that the initial entrainment of the clock may have varied across the experiment, and the different light/dark input for each experiment may have caused phase changes. However, the hierarchical approach can be used for scenarios where there is reason to assume that switch-times are similar between experiments. Inference is achieved by formulating an appropriate MCMC algorithm that samples from $p(\Theta_\delta, \mathbf{Y})$ (Gamerman and Lopes, 2006). In practice, we specified $p(\delta^{(i)}|\Theta_\delta)$ by a Gamma distribution parameterized to have mean μ_δ and variance σ_δ^2. We assigned a Gamma prior distribution to the mean μ_δ and an exponential distribution to the coefficient of variation (CV) σ_δ/μ_δ. The CV is used, as the mean value is close to 0; therefore, it is a more robust parameter for sampling σ_δ^2. We derive informative priors from degradation rates from >13 K *A.thaliana* genes from Narsai *et al.* (2007):

$$\delta^{(i)} \sim \gamma(\mu_\delta, \sigma_\delta^2);\ \mu_\delta \sim \gamma(\mu_{Nar}, \sigma_{Nar}^2);\ \frac{\sigma_\delta}{\mu_\delta} \sim \mathrm{Exp}(\lambda_{Nar})$$

where μ_{Nar} and σ_{Nar}^2 are estimated from the mean degradation rates in the dataset, and λ_{Nar} is taken to be the 95% percentile of the CV distribution from the data. All models use the same informative population-level prior distributions, rather than the gene-specific estimates because of large potential differences between experimental protocols. The approach taken by Narsai *et al.* was to fit exponential decay least squares regression models to microarray data collected at six time points after the transcriptional inhibitor actinomycin D was added. Although this is conceptually a straightforward approach for estimating degradation rates, it is neither clear whether the method completely inhibits transcription nor whether its invasiveness has an effect on the degradation processes themselves. Therefore, a strong gene-specific prior derived from the Narsai *et al.* estimates could not be justified.

Figure 4 shows all estimates of the mRNA degradation rate of the core clock gene LHY from Narsai *et al.*, the hierarchical model and, for comparison, the non-hierarchical (independent) model. The Narsai *et al.* estimate has an approximate range in half-life of 1.5–2.25 h, and our posterior estimates are broadly in a similar range. The estimated joint distribution summarizes the variability of the three experiments and provides a theoretically rigorous summary statistic of the degradation rate (which cannot be achieved by averaging over the independent results). Estimates from the individual models show a range of estimates from 1.3 to >2 h. Results for E2 and E3 are more variable probably because they cover shorter timescales of 17.5 and 6 h, whereas E1 covering two circadian cycles provides more precise estimates despite smaller sampling frequency. An interesting observation is the difference between the E1 estimate and E2, E3 and Narsai *et al.* estimates. There are a number of potential reasons for the difference, given the experiments were performed over different time intervals and in different laboratories. However, it may also be related to the light conditions, as E1 is the only experiment incorporating two 8 h dark periods. Light is a key driver in the *A.thaliana* circadian clock, and a recent study has suggested a light-specific degradation rate for CCA1, a core partner of LHY (Yakir *et al.*, 2007), and a current model for the *A.thaliana* clock uses different degradation rates for the LHY/CCA1 component in light and dark (Pokhilko *et al.*, 2012). This result complements the hypothesis that light plays a role in mRNA degradation in at least some core circadian clock components. A possible extension to the model could include light/dark cycles, for which separate degradation rates are assumed.

For further comparison with the Narsai *et al.* estimates, we grouped all genes for which there was an estimate available in their dataset (136 of the 200 genes) into five broad mRNA

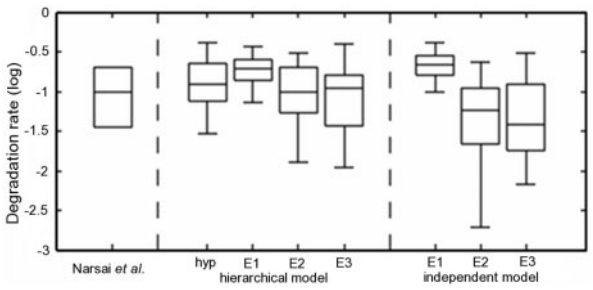

Fig. 4. Box plots of degradation rate estimates for gene LHY. Degradation rate estimates visualized for the gene LHY obtained experimentally (left), hierarchical model; hyperdistribution and the three individual experiment estimates (centre) and estimates from the independent models for each individual experiment (right). Data for experimentally obtained degradation rate estimates are taken from Narsai *et al*. (2007). Box plots for model estimates show 25th–75th percentiles and whiskers at 5th and 95th percentiles, whereas the Narsai *et al*. box plot shows mean ±2 standard error

stability groups based on half-life, as used by Narsai *et al*. (0–1, 1–3, 3–6, 6–12 and >12 h). We find that 32% (43 of 136 genes) have a similar hyperdistribution estimate using our approach to the Narsai *et al*. study, and agreement between the three individual estimates and Narsai *et al*. ranges from 34% up to 36% (Supplementary Table S8). However, despite this overlap, there is also considerable variability in degradation rates between the experiments, which may be natural variability or because of the experiments carried out under different conditions.

4 DISCUSSION

The aim of this article is to present a novel approach for identifying timing of transcriptional activity from time series mRNA expression data. The model introduced here consists of a piecewise linear simple ODE model of mRNA dynamics, which can be fitted efficiently with a RJMCMC sampler to estimate gene-specific parameters, i.e. mRNA stability and number and times of switches in transcriptional activity. Estimation and performance of the algorithm is investigated for synthetic data of varying sampling frequencies and noise levels in a simulation study. With the example of time series microarray data from 200 circadian genes, further directions are explored exploiting different aspects of the model output. Namely, using the timing of the switches as a basis for clustering, which, when combined with promoter motif data, seems to identify more significant groups of motifs than simple profile clustering with promoter motif data, potentially implying a stronger correlation with regulatory mechanisms. We also explore the potential for the estimation of mRNA degradation rates. Usually, degradation rate studies involve treatment with a transcriptional inhibitor, such as actinomycin D, or translational inhibitor, such as cycloheximide. It is not clear whether such inhibition is ever achieved fully and whether such treatments have undesired side-effects on degradation, and may, therefore, impact on estimated rates in unpredictable ways. The model introduced in this study has several advantages over a transcription inhibition study. The primary advantage is that a specific experiment does not have to be designed and performed, often at great cost in time and resources, to produce a suitable dataset for degradation estimation, effectively allowing recycling of existing datasets further increasing their potential scientific value. As only free-running mRNA expression dynamics are required, potential side-effects introduced by using a chemical inhibitor can be avoided. We demonstrate how to pool data from several experiments in a theoretically rigorous way with a Bayesian hierarchical model. Degradation estimates can easily be obtained for suitably resolved time series and can be compared between different experimental conditions. As the number of large high-resolution gene expression time series datasets publicly available is likely to increase with the development of cheaper and faster high-throughput technologies, new methods are required to analyse these data. The model proposed here is mechanistic yet is flexible and rich enough to capture a wide range of expression dynamics observed in mRNA time series data, from steady-state behaviour to oscillatory expression. At the same time, it is simple enough to be estimated with feasible computational time for thousands of genes. Using a mechanistic model to identify transcriptional switch points is likely to strongly contribute to efforts in elucidating and understanding regulatory interactions within transcriptional networks.

ACKNOWLEDGEMENTS

This research is part of the PRESTA (Plant Responses to Environmental STress in *Arabidopsis*) project, and the authors thank the group for high-resolution *A.thaliana* microarray and motif data and for helpful discussions and advice. They thank M. Costa, K. Hey, S. Veflingstad and D. Woodcock for discussions on methodology.

Funding: Biotechnology and Biological Sciences Research Council (BB/F005806/1); Engineering and Physical Sciences Research Council (EP/C544587/1 to D.A.R.).

Conflict of Interest: none declared.

REFERENCES

Aase,S.O. and Ruoff,P. (2008) Semi-algebraic optimization of temperature compensation in a general switch-type negative feedback model of circadian clocks. *J. Math. Biol.*, **56**, 279–292.

Androulakis,I.P. *et al.* (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annu. Rev. Biomed. Eng.*, **9**, 205–228.

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.*, **10**, 341–356.

Baxter,L. *et al.* (2012) Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell*, **24**, 3949–3965.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Calvano,S.E. *et al.* (2005) A network-based analysis of systemic inflammation in humans. *Nature*, **437**, 1032–1037.

Chechik,G. and Koller,D. (2009) Timing properties of gene expression responses to environmental changes. *J. Comput. Biol.*, **16**, 279–290.

Cooper,S.J. *et al.* (2006) Comprehensive analysis of transcriptional promoter structure and function of 1% of the human genome. *Genome Res.*, **16**, 1–10.

Denison,D.G.T. *et al.* (1998) Automatic Bayesian curve fitting. *J. R. Stat. Soc. B*, **60**, 333–350.

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

Finkenstädt,B. *et al.* (2008) Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, **24**, 2901–2907.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gamerman,D. and Lopes,H.F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd edn. Chapman & Hall/CRC, Boca Raton.

Golightly,A. and Wilkinson,D.J. (2011) Bayesian parameter inference for stochastic biochemical network models using particle MCMC. *Interface Focus*, **1**, 807–820.

Green,P.J. (1995) Reversible jump Markov chain Monte Carlo computations and Bayesian model determination. *Biometrika*, **82**, 711–732.

Grün,B. *et al.* (2012) Modelling time course gene expression data with finite mixtures of linear additive models. *Bioinformatics*, **28**, 222–228.

Harper,C.V. *et al.* (2011) Dynamic analysis of stochastic transcription cycles. *PLoS Biol.*, **9**, e1000607.

Heard,N.A. *et al.* (2005) Bayesian coclustering of anopheles gene expression time series: Study of immune defense response to multiple experimental challenges. *Proc. Natl Acad. Sci. USA*, **102**, 16939–16944.

Higo,K. *et al.* (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res.*, **27**, 297–300.

Kiddle,S.J. *et al.* (2010) Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*. *Bioinformatics*, **26**, 355–362.

Komorowski,M. *et al.* (2009) Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, **10**, 343.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.

Lamesch,P. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Narsai,R. *et al.* (2007) Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. *Plant Cell*, **19**, 3418–3436.

Opper,M. and Sanguinetti,G. (2010) Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics*, **26**, 1623–1629.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.

Pokhilko,A. *et al.* (2012) The clock gene circuit in *Arabidopsis* includes a repressilator with additional feedback loops. *Mol. Syst. Biol.*, **8**, 574.

Sanguinetti,G. *et al.* (2009) Switching regulatory models of cellular stress response. *Bioinformatics*, **25**, 1280–1286.

Schaffer,R. *et al.* (1998) The *late elongated hypocotyl* mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. *Cell*, **93**, 1219–1229.

Schliep,A. *et al.* (2004) Robust inference of groups in gene expression time-courses using mixtures of HMMs. *Bioinformatics*, **20** (**Suppl. 1**), i283–i289.

Sclep,G. *et al.* (2007) CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. *BMC Bioinformatics*, **8**, 400.

Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.

Windram,O. *et al.* (2012) Arabidopsis defense against *Botrytis cinerea*: chronology and regulation deciphered by high-resolution temporal transcriptomic analysis. *Plant Cell*, **24**, 3530–3557.

Yakir,E. *et al.* (2007) CIRCADIAN CLOCK ASSOCIATED1 transcript stability and the entrainment of the circadian clock in Arabidopsis. *Plant Physiol.*, **145**, 925–932.

Yoneya,T. and Mamitsuka,H. (2007) A hidden Markov model-based approach for identifying timing differences in gene expression under different experimental factors. *Bioinformatics*, **23**, 842–849.