Reconstruction of transcriptional dynamics from gene reporter data using differential equations

Citation for published version:

Digital Object Identifier (DOI):
10.1093/bioinformatics/btn562

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Bioinformatics

Publisher Rights Statement:
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Systems biology

Reconstruction of transcriptional dynamics from gene reporter data using differential equations

Bärbel Finkenstädt¹,*, Elizabeth A. Heron¹,², Michal Komorowski¹,², Kieron Edwards³, Sanyi Tang²,†, Claire V. Harper¹, Julian R. E. Davis⁵, Michael R. H. White⁴, Andrew J. Millar³ and David A. Rand²

¹Department of Statistics, University of Warwick, Coventry, CV4 7AL, ²Systems Biology Centre, University of Warwick, Coventry, CV4 7AL, ³Institute for Molecular Plant Sciences, University of Edinburgh, Edinburgh EH9 3JH, ⁴Department of Biology, University of Liverpool and ⁵School of Medicine, University of Manchester, Liverpool, Manchester

Received on July 24, 2008; revised on October 1, 2008; accepted on October 25, 2008

Advance Access publication October 30, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Promoter-driven reporter genes, notably luciferase and green fluorescent protein, provide a tool for the generation of a vast array of time-course data sets from living cells and organisms. The aim of this study is to introduce a modeling framework based on stochastic differential equations (SDEs) and ordinary differential equations (ODEs) that addresses the problem of reconstructing transcription time-course profiles and associated degradation rates. The dynamical model is embedded into a Bayesian framework and inference is performed using Markov chain Monte Carlo algorithms.

Results: We present three case studies where the methodology is used to reconstruct unobserved transcription profiles and to estimate associated degradation rates. We discuss advantages and limits of fitting either SDEs ODEs and address the problem of parameter identifiability when model variables are unobserved. We also suggest functional forms, such as on/off switches and stimulus response functions to model transcriptional dynamics and present results of fitting these to experimental data.

Contact: b.f.finkenstadt@warwick.ac.uk

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Imaging data from luciferase (LUC) and green fluorescent protein (GFP) reporters combined with fluorescent tagging of protein can provide very high quality data with good temporal resolution (Millar et al., 1995; Nelson et al., 2004). In this case the actual imaging time series is approximately proportional to the abundance of an artificial protein. The underlying transcriptional dynamics are unobserved and are masked by two degradation processes, namely of reporter mRNA and reporter protein. In this study, we address the problem of back-calculating from the observed protein activity to the hidden transcriptional dynamics where it is of interest to estimate the associated rates of degradation as part of the analysis.

We formulate a probability model based on (stochastic) differential equations which provides the mechanistic rules for the back-calculation. In practise heterogeneous datasets may be available from different experiments which contain information about the transcription process and model parameters. Data sources may be of different quality and time resolution, as well as from single cells or an aggregated population of cells. Longitudinal measurements are discrete in time and can be irregularly spaced or on different time scales for different variables. Other realistic shortcomings of the data are that time-course measurements may not correspond to the same biological sample, or data on different variables may not be matched in time which would be preferable for fitting a multivariate dynamical model. As the quality and quantity of such datasets supports more or less complex modeling approaches, we consider both stochastic differential equations (SDEs) and ordinary differential equations (ODEs) with measurement noise. Information on rate constants may be incorporated through prior distributions in a Bayesian approach. We first describe the models and the statistical methods used for its inference. Then, we present three case studies each with the aim of reconstructing transcription and inferring any identifiable degradation rates from reporter gene data using available heterogeneous sources of data. These case studies serve to demonstrate the adaption of the methodology to different experimental scenarios.

2 MODELS AND INFERENCE

It is now well understood that, because of the stochastic nature of reaction events and the presence of internal noise due to the fluctuations in the molecular environment of the cell, regulatory and signalling systems are intrinsically stochastic. To develop a stochastic model one can attempt to model the individual stochastic events involved, such as binding of the transcription factors, the assembly and initiation of the polymerase and transcription. Although an exact simulation algorithm of the corresponding stochastic processes is provided by Gillespie (1977, 1992) such models are too detailed for there to be any hope of fitting to current data with its limitations. SDEs provide a good approximation of molecular population systems when one can assume that there is a
We now introduce the two approaches and their likelihood derivation. This approach is rigorously modeling the volatility of the stochastic process whilst the modeler makes assumptions about the residual process and measurement error. If the error process is assumed independent then the likelihood in the mean ODE approach is

\[ L_{\text{ODE}}(\theta; Y) = \prod_{i=1}^{T} \Phi(y_{i+1} - y_i; \delta \tau_i, \Sigma \delta \tau_i) \]

where \( \Phi(x; \mu, \Sigma) \) denotes the bivariate normal density with mean vector \( \mu \) and variance matrix \( \Sigma \). Justifications for this approximation are given in Kloeden and Platen (1999).

Mean ODE approach: suppose there is a solution path \( \mu(t; \theta) = (M(t), P(t); \theta) \) to the system in (1) from unknown initial conditions \( \Delta t = t_{i+1} - t_i \). Then a natural probabilistic model is to assume that \( Y \) has a joint distribution with mean function \( \mu(t; \theta) \) and a variance function \( \sigma^2(t; \theta) \). The distribution function and variance are specified according to assumptions that the modeler makes about the residual process and measurement error. If the error process is independent then the likelihood in the mean ODE approach is

\[ L_{\text{ODE}}(\theta; Y) = \prod_{i=1}^{T} \phi(y_i; \mu(t_i), \sigma^2(t_i), \theta), \]

The Wiener process, or Brownian motion, is a continuous time stochastic process that has independent normally distributed increments.
where θ now incorporates initial conditions (M₀, P₀) and g is a suitably chosen probability distribution.

Inference: by Bayes’ theorem the posterior distribution is

$$\pi(\theta | Y) \propto L(\theta | Y) \pi(\theta), \quad (6)$$

where L is the likelihood function, derived for either the ODE or SDE approach, and π(θ) are prior densities of model parameters. Sampling from the posterior distribution is usually achieved using Markov chain Monte Carlo (MCMC), where each element of θ is updated by using an appropriately constructed Metropolis–Hastings acceptance/rejection scheme based on either random walk or independence proposals (Gamerman and Lopes, 2006). The reason for choosing a Bayesian approach combined with a MCMC algorithm is 2-fold: first, the Bayesian methodology is flexible allowing for portability of inference results between different experimental studies in a well-defined way and this is highly relevant to studies in systems biology. Second, the probabilistic imputation of missing data and/or unobserved variables can be implemented in a straightforward way as part of an MCMC sampler.

Discrete data and unobserved variables: molecular time-series data are discretely measured and it cannot be guaranteed that the sampling interval is small enough for the approximation (*) to work well. A remedy suggested in econometric applications of SDEs (Durham and Gallant, 2002; Elerian et al., 2001) is to augment the observed data by introducing a number of latent or unobserved data points, called a bridge, in-between the measurements with the aim of creating a virtual fine discrete time grid for which the assumption in (*) is valid. The bridges are treated as missing or latent data. Let Y* denote the collection of all latent data. We wish to sample from the joint distribution of all latent data. We wish to sample from the joint of all latent data. We wish to sample from the joint of all latent data. We wish to sample from the joint of all latent data.

Inference: by Bayes’ theorem the posterior distribution is

$$\pi(\theta, Y^* | Y) \propto L(Y^*, Y | \theta) \pi(\theta) \pi(Y^* | \theta), \quad (7)$$

where $$L(Y^*, Y | \theta)$$ is the approximated augmented likelihood. This is achieved by sampling in turn from the full-conditionals densities of θ | Y*, Y and Y* | θ, Y (Tanner and Wong, 1987). Thus, in the framework of an MCMC, one can generate proposal bridge processes and accept these with an appropriately constructed acceptance probability. In practise, we have used (Heron et al., 2007) a bridging method based on an independent sampler suggested by Elerian et al. (2001) (see Supplementary Material). The treatment of other forms of missing data, such as unobserved variables as part of the inference algorithm is theoretically the same. In practise, this is challenging as the dimension of the posterior density in (7) can become very large. We present applications of bridge building and stochastic reconstruction of unobserved processes in our case studies. One also needs to decide upon the size of a virtual sampling interval for which one can safely assume that (*) holds. Since there are no analytical solutions we base our choice on Monte Carlo studies of simulated systems.

3 CASE STUDIES

3.1 Case study 1: red light pulse experiment

The Arabidopsis thaliana gene Chlorophyll A/B binding Protein 2 CAB2 is regulated by light and the circadian clock (Millar and Kay, 1996). The aim here is to estimate degradation rate of CAB2 mRNA and to reconstruct the transcriptional dynamics of the CAB2:LUC reporter gene as a result of a 20 min red-light induction. At subjective dawn on the 6th day of the experiment (see Supplementary Material for a description of experiment), the grown Arabidopsis seedlings were given a 20 min red light pulse to induce CAB2 expression. Samples were harvested at the indicated time points and total RNA and total protein was extracted. Steady-state levels of LUC mRNA were measured by quantitative PCR (Q-PCR) and an in vitro LUC assay was used to measure LUC activity in the protein samples. Concurrently, red light-pulsed seedlings were also imaged for LUC activity using light sensitive cameras (Millar et al., 1995). This allows the measurement of LUC activity within the same seedlings throughout the entire experiment, whereas the in vitro LUC assays and Q-PCR experiments necessarily sacrificed different samples for each time point. All data are probes from whole leaves (plots of all time series in Supplementary Material) representing cell populations and the activity of the clock gene can be assumed to be synchronized between cells by the light pulse. There are three replicates of each measurement variable sampled every half hour for a length of 7 h. Matching control replicates that have not been subject to light induction were sampled for the same time length albeit more sparsely for the Q-PCR and in vitro assay data.

Assuming that molecular populations all scale differently with the Q-PCR, in vitro and in vivo imaging data, we use (1) to describe the dynamics of mRNA and imaged LUC protein and add a third equation

$$dPv/dt = apR. M(t) - dpPv, \quad (8)$$

which represents the protein dynamics measured by the in vitro LUC protein assays (see Supplementary Material for full model statement). The two protein equations are identical except for differently scaled translation rates ap and ar. Furthermore, a constant cr is added to the imaging data to represent some threshold level at which the camera is able to detect a signal. To specify a form for the transcription τ(t) consider an indicator function L(t)=1 for the time of the red light pulse, and L(t)=0 otherwise [L(t)=0 for all control experiments]. The response of mRNA transcription to the stimulus can then be modeled as a convolution of L(t) and d(u) which is a probability density for the waiting time u between the pulse and the initiation of transcription, i.e.

$$\tau(t) = a_0 \left( \int_0^\infty d(u) L(t-u) du + \tau \right). \quad (9)$$

where τ represents a baseline transcription. We take d(u) to be a Gamma density with mean µu and SD σu to be estimated. The specification in (9) is motivated by the fact that it successfully reproduced the qualitative features observed in the data in preliminary model simulations and because d is flexible. Since data are from aggregated cell populations, the imaged protein data are very smooth and successive data points of the Q-PCR and in vitro time series come from different samples of cell populations, we choose to fit the model using the mean ODE approach with independent error. To ensure all variables are strictly non-negative, we used an independent Gamma distribution for g in the likelihood (5) for each of the three variables where parameters were specified to have mean process equal to an ODE solution and time constant variance $$\sigma^2_M, \sigma^2_P, \sigma^2_C$$. Applying (5) the likelihood of replicate

$$r = 1, 2, 3$$

is

$$L(\theta' | Y^*) = \prod_{i=1}^{T^R} g(y_i^R | \mu(t_i), \theta') \prod_{j=1}^{T^C} g(y_j^{gC} | \mu(t_{j}), \theta''), \quad (10)$$

where $$y_i^R$$ is the vector of observed data points $$i = 1, ..., T^R$$ for variables $$M, P, P$$, for replicate r under the red light experiment, $$y_j^{gC}$$ denotes observed data points $$j = 1, ..., T^C$$ for the corresponding control experiment and g is a product of Gamma densities. The ODE model was fitted to each of the replicates $$r = 1, 2, 3$$ and to the average of the replicates where prior distributions for all parameters were chosen to be uninformative. Results of posterior estimates are summarized in Table 1 and the model fit can be seen in Figure 1. The mean delay time between light induction and transcription is about 2 h with almost all transcription happening between 0.8 h and 3.2 h after...
Table 1. Case 1: posterior results for selected parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Average</th>
<th>r1</th>
<th>r2</th>
<th>r3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \delta_M )</td>
<td>1.542 (0.019)</td>
<td>1.726 (0.044)</td>
<td>1.417 (0.121)</td>
<td>3.526 (0.315)</td>
</tr>
<tr>
<td>half-life</td>
<td>0.45 h</td>
<td>0.4 h</td>
<td>0.49 h</td>
<td>0.2 h</td>
</tr>
<tr>
<td>( \mu_T )</td>
<td>2.008 (0.011)</td>
<td>2.101 (0.014)</td>
<td>1.902 (0.045)</td>
<td>2.362 (0.0289)</td>
</tr>
<tr>
<td>( \sigma_T )</td>
<td>0.631 (0.013)</td>
<td>0.692 (0.014)</td>
<td>0.686 (0.039)</td>
<td>0.723 (0.0217)</td>
</tr>
<tr>
<td>( \tau )</td>
<td>0.012 (0.001)</td>
<td>0.014 (0.001)</td>
<td>0.014 (0.002)</td>
<td>0.013 (0.002)</td>
</tr>
<tr>
<td>( \delta_P )</td>
<td>0.305 (0.0045)</td>
<td>0.286 (0.0040)</td>
<td>0.272 (0.010)</td>
<td>0.365 (0.0093)</td>
</tr>
<tr>
<td>half-life</td>
<td>2.27 h</td>
<td>2.42 h</td>
<td>2.5 h</td>
<td>1.9 h</td>
</tr>
</tbody>
</table>

Posterior means and SDs of selected estimated parameters (See Supplementary Material for all parameters), where the red light pulse model was fitted to average data and to single replicate datasets denoted by r1, r2, r3. Estimated rates are per hour. Degradation rates are translated into half-lives as follows: half-life (in hours)=ln(2)/degradation rate.

ON/OFF switching transcription profile from the following two experimental datasets:

1. Native mRNA Q-PCR data: Q-PCR measurements were taken at 2 h intervals over 72 h on CCA1 mRNA entrained under a photoperiod of 18 h before being released into constant light. The data used are an average of concentrations relative to the start of two biological replicates.

2. Protein imaging: high resolution imaging data for a different experiment with identical conditions as for data (1) were sampled at 1.5 h intervals over a length of 91.5 h on LUC protein activity resulting from LUC reporter constructs fused to the CCA1 promoter. Similar to case study 1, all data come from whole leaves and thus represent a population of cells where the activity of the clock gene is synchronized between cells during the exposure to dark. light cycles during the entrainment period (see Supplementary Material for further details of experiment). The data used are an average of concentrations relative to the start of 20 replicates.²

No data were available for the CCA1:LUC mRNA. However, if we assume that CCA1:LUC and CCA1 mRNA have the same transcriptional dynamics, then the available two time series are connected in a dynamic model with three variables where LUC mRNA and LUC protein dynamics are described by (1) and a further equation

\[ \frac{dM_{LUC}}{dt} = \tau(t) - \delta_M M_{LUC}(t) \]

is added for the native CCA1 mRNA. We assume that observed variables are proportional to \( M_t \) and \( P_t \) populations with scaling factors \( \delta_M \) and \( \delta_P \), while \( M_t \) is unobserved. To describe the oscillatory nature of the data, we consider an ON/OFF switching function for the transcription \( \tau(t) = \tau_{OFF} \) if transcription is active at time \( t \), and \( \tau(t) = \tau_{ON} \) if transcription is inactive. This function has the advantage of being interpretable and parsimonious. If it produces realistic oscillations then its simple structure makes it an interesting ingredient to models of larger networks. Let \( S = (s_1, s_2, ..., s_k) \) where \( s_1 < s_2 < ... < s_k \) are the times at which a switching between an ON and OFF state occurs. They are estimated as part of the MCMC algorithm where we assume that the number of switches and the initial state are known.³ To set the phase of the clock both the data series experienced a light–dark (LD) cycle of 18 h of L and 6 h of D at the beginning of the sampling period and this seems to generate a higher amplitude. We allow for this by setting the transcription on-rate to \( \frac{\tau_{OFF}}{\tau_{ON}} \) during the first 35 h (allowing also for some delayed effect of the dark period). For purpose of estimation, the mean ODE approach will be appropriate for similar reasons as case study 1. However, an SDE approach is a superior theoretical model that should be considered even if data do not (yet) strictly comply with its underlying assumptions. We use this case study to show the application of both approaches.

SDE approach: consider a system of SDEs formulated analogously to (2). Since \( M_t \) is unobserved it can be imputed stochastically as realizations of the SDE but the cost of computation is high. Simulation studies suggested that the more practicable way of imputing \( M_t \) as solution to an ODE from an initial condition \( M_0 \) to be estimated had no discernable impact on our inference results here. In order to fit an SDE model to discrete data points for \( M_t \) and \( P_t \), we augment the coarse grid to a virtually fine grid [for which assumption (*) is valid] by imputing auxiliary data in the form of bridges. Let \( \theta = (S, \tau_{OFF}, \delta_M, M_0, \delta_P, M, \alpha, \delta_P, S) \) denote the vector of unknown parameters and let \( M_{S}^\alpha \) and \( P_{S}^\alpha \) be the auxiliary data for \( M_t \) and \( P_t \), respectively.

³For computational precision, we amplified the mRNA concentrations by factor 10² and the protein concentrations by 10³.

³The number of switches and initial state are fairly obvious here. The inference algorithm can, however, be generalized to allow for an arbitrary number of switches and where the initial state is estimated. We will describe work on this elsewhere.
Table 2. Case 2: posterior results for selected parameters

<table>
<thead>
<tr>
<th></th>
<th>(\delta M_g)</th>
<th>(\delta M)</th>
<th>(\delta P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDE</td>
<td>0.426 (0.0043)</td>
<td>1.54 (0.019)</td>
<td>0.072 (0.0057)</td>
</tr>
<tr>
<td>ODE</td>
<td>0.313 (0.0273)</td>
<td>1.42 (0.101)</td>
<td>0.075 (0.0018)</td>
</tr>
</tbody>
</table>

Posterior mean and standard error estimates of selected parameters of model in case 2 using the SDE and mean ODE approach. All rates are per hour. Estimates for all parameters and switch times are provided in Supplementary Material.

Then according to (7) the posterior distribution for the unknown \(\Theta, M_g^*, P^*\) is given by

\[
\pi(\theta, M_g^*, P^*| M_g, P) \propto L(M_g, P, M_g^*, P^*| \theta) \pi(\theta),
\]

where we approximate \(L(M_g, P, M_g^*, P^*| \theta)\) with the augmented likelihood in (4) for small sampling intervals for all observed and auxiliary data, i.e. \(y = (M_g, P, M_g^*, P^*)\). More details of the SDE inference algorithm are provided in the Supplementary Material.

Mean ODE approach: here the likelihood is given by (5) where the unobserved variable \(M\) is reconstructed as an solution of an ODE from an initial condition \(M_0\) to be estimated. The density \(g\) was specified to be the product of two independent normal distributions with mean equal to the joint ODE solutions for \(M_g\) and \(P\) and with variance parameters \(\sigma_{M_g}^2\) and \(\sigma_P^2\). We have set \(\tau_{off} = 0\) for the off-time as initial estimations showed that it was not different from zero. As the variables are concentrations relative to initial conditions the ODE solutions are assumed to start at one. Thus, the parameter vector for the mean ODE approach is \(\theta = (S\mu, \gamma_0, \tau_{off}, \delta M_g, \delta M, \alpha, \beta, \sigma_{M_g}, \sigma_P)\).

To ensure identifiability in both estimation approaches the prior distribution for \(CCA1:LUC\) mRNA degradation \(\delta M\) has to be informative. We hence used a Gamma distribution with mean 1.542 and SD 0.019, corresponding to the results in Table 1. All other priors were taken independently uniform in an attempt to estimate all remaining parameters only from the experimental data at hand. Posterior estimates are given in Table 2. Figure 2 shows the transcription profile and model fits for both approaches. The plots suggest that the switch model is remarkably able at reproducing the observed oscillations. The main feature of the reconstructed profiles is that the inactive times (around 15–18 h) are at least twice as long as the active times (around 7 h) and this produces the pronounced asymmetric cycles in the protein and mRNA time series. The estimates also suggest that there is a shorter but larger burst of transcription during the dark period. Both approaches deliver similar posterior rates for degradation. Our results for \(CCA1\) mRNA degradation are in remarkable agreement with the analysis in Yakir et al. (2007) whose estimates correspond to 0.23 in darkness to 0.46 in light for \(\delta M_g\). Both approaches reliably estimate the half-life of the LUC protein to be around 9.5 h. This is surprisingly long and is probably due to a lack in provision of luciferin. The most notable difference between the two approaches lies in the variance estimation. The SDE approach has to deal with the estimation of the two scaling parameters, \(\delta P\) and \(\alpha M_g\). We find that their identification from the experimental data is problematic as convergence could not achieved although this did not affect convergence of all other parameters. The two scaling parameters were thus sampled within some chosen bounded region of parameter space. In particular, in order for the bridge sampling to remain numerically stable for low values of the mRNA series, the sampling of \(\delta P\) had to be bounded to artificially low values. The identifiability problem of the scaling parameters leads to problems in realistically quantifying the volatility. The estimated intervals in Figure 2 illustrate this for the mRNA series. For the mean ODE approach variability is measured by the posterior standard error of the fit similar to a regression and the graph shows that predictions can be made more precisely about the protein dynamics than about the native mRNA. This is reflecting the fact that the protein data are more aggregated and smoother time series than the mRNA series.

3.3 Case study 3: Stochastic transcription for single cell data

In this experiment protein activity was imaged from GH3 rat pituitary cells stably transfected with a construct comprising a 5 kb human prolactin gene promoter fragment linked to a destabilized EGFP reporter gene (hPRL-d2EGFP) (see Supplementary Material for details of experiment). Images were taken 108 times in 15 min intervals giving a total of 27 h of data for a single cell (Fig. 3). We assume that the dynamics are described by the SDE model in (2). Since \(M\) is not observed we cannot identify the degradation rates \(\delta M, \delta P\) and a strongly informative prior density is needed. Here, we assume that each of them have an independent Gamma distribution with mean 0.4 for \(\delta M\) and 0.5 for \(\delta P\). The prior variance was arbitrarily chosen to be small at 0.02 for both parameters. Since \(M\) is unobserved we can arbitrarily fix \(s\) to 1. Given the particular form of an experiment, where transcription is induced and afterwards comes back to its initial level, we have specified \(\tau(t)\) as follows

\[
\tau(t) = \begin{cases} 
 b_0 \exp\left(\frac{(t-b_3)^2}{b_1}\right) + b_4 & t \leq b_3 \\
 b_0 \exp\left(\frac{(t-b_2)^2}{b_1}\right) + b_4 & t > b_3,
\end{cases}
\]

where the parameters \(b_j\) are to be estimated. Priors for parameters different than degradation rates were intended to be uninformative. Here, we used exponential priors with means given in Table 3. The challenge for inference here is to integrate over a fully unobserved process \(M\) whilst sampling bridges

\[^{4}\text{We could not set } \tau_{off} = 0 \text{ in the SDE case for the practical problem that the bridge building algorithm becomes numerically unstable for values of the mRNA too close to zero.}\]

\[^{5}\text{These rates were motivated by preliminary estimation using a small dataset from other experiments. They are used here only to demonstrate the case as their estimates may change if more data were available.}\]
Table 3. Case 3: posterior inference results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Prior</th>
<th>Simulation</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_M$</td>
<td>0.44</td>
<td>$\Gamma(0.44,0.02)$</td>
<td>0.56 (0.36–0.92)</td>
<td>0.45 (0.26–0.82)</td>
</tr>
<tr>
<td>$\delta_P$</td>
<td>0.52</td>
<td>$\Gamma(0.52,0.02)$</td>
<td>0.59 (0.38–0.89)</td>
<td>0.71 (0.45–1.09)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>20</td>
<td>$\exp(100)$</td>
<td>16.97 (6.54–78.98)</td>
<td>0.46 (0.14–1.51)</td>
</tr>
<tr>
<td>$\sigma_P$</td>
<td>0.2</td>
<td>$\exp(1)$</td>
<td>0.17 (0.09–0.3)</td>
<td>2.11 (1.24–3.56)</td>
</tr>
</tbody>
</table>

Parameter values used in simulation study. Priors, posterior medians and 95% credibility intervals inferred from both simulated and experimental data. Rates are per hour. $\Gamma(\mu,\sigma^2)$ denotes Gamma distribution with mean $\mu$ and variance $\sigma^2$. Full list of all parameter estimates is provided in Supplementary Material.

to augment the discretely observed $P$. Let $P^*$ denote the vector of bridges augmenting the $P$ process and $M^*$ denote the latent $M$ variable [we chose a grid size of 1 min for which we assume that (*) holds]. The vector of unknown parameters is $\theta = (\delta_M, \delta_P, \alpha, \sigma_P, b_0, b_1, b_2, b_3, b_4)$. The posterior distribution takes the form

$$
\pi(\Theta, M^*, P^*| P) \propto L(M^*, P^*| \Theta) \pi(\Theta)
$$

where we approximate $LM^*, P^*, P| \Theta$ with the likelihood (4) for the augmented data case, i.e. $\gamma = (M^*, P^*, P)$. In practice, this is a challenging sampling problem as the dimension of the posterior is very large and traces were highly autocorrelated. Faster convergence is achieved by reparameterizing the model (details of this and the algorithm are given in the Supplementary Material). The algorithm was first tested on simulated data from the SDE model with chosen parameters (Table 3). Artificial data are simulated on a fine scale of 15/51 min and coarse data are extracted for $P$ at 15 min intervals. The simulated and observed time series, and the reconstructed $r(t)$ are shown in Figure 3. Posterior inference results are given in Table 3. Note that since $M$ is not scaled the transcription profile corresponds to molecular population sizes which here are about 150 mRNA molecules per hour. This case study demonstrates that for high frequency single cell data the SDE approach can be extremely powerful as it allows estimation of absolute transcription rates in terms of molecule numbers and since $\sigma_P$ can be estimated it is possible to calculate back to molecular levels of protein and translation rate. The need for precise prior information about degradation rates is irrespective of either SDE or ODE approach. The problem of non-identifiability of these parameters is due to not observing $M$ as one can infer both degradation rates in either approach if both $M$ and $P$ are observed.

4 DISCUSSION

In this study, we suggest a dynamical model relating protein and corresponding mRNA dynamics via transcription and translation and suggest methods for model fitting. The applications here were motivated by the availability of gene reporter data but the model and methodology apply to many other scenarios where it is of interest to link protein and mRNA dynamics. While a stochastic model, such as (2) applies to single cell data, caution needs to be exercised in formulating an ODE model, such as (1) for multi-cell data. In order to reasonably assume such a joint mechanistic model it is essential that the individual cell activities are synchronized with respect to the gene of interest. Rate constants associated with processes of degradation, transcription and translation arise as model parameters and it is an important question whether these can be identified. In addition to a functional kind of non-identifiability of parameters in complex dynamic models as considered in Hengl et al. (2007) here, we find that practical or statistical non-identifiability of model parameters may result from unobserved variables. Case study 1 demonstrates that one can estimate all rate constants in systems of equations of the type given in (1) if all model variables—albeit coarse—are observed over time. Inference precision increases with the frequency at which the processes are sampled. In contrast, Cases 2 and 3 have latent variables and model inference is only feasible with informative prior knowledge of some parameters. Simulation studies of the model (using artificial parameters) help in identifying which sets of parameters need to be informed from other experiments. In case 3, prior knowledge of both degradation rates was needed as with $M$ unobserved, parameters can trade-off giving rise to protein dynamics that is virtually indistinguishable via likelihood from the observed protein process. The specification of the functional form for the transcription profile also plays a role in practical identification. Even if $M$ is observed the parameter estimates associated with transcription and degradation are correlated for obvious reasons. Such correlations affect precision of estimates and convergence of the Markov chain but can be alleviated by sampling more frequently, choosing a parsimonious functional form for transcription, and by technical aids such as the construction of independence samplers and reparameterization of the model. We believe that the functional specifications for $r(t)$ suggested in our case studies are useful in conjunction with gene transcription. A theoretical application of the switch function in clock modeling can be found in Aase and Ruoff (2008). Although the estimation of the switch model seems too high dimensional for datasets with many switches, this could be overcome by assigning probability distributions to the on- and off-times in the framework of a Bayesian hierarchical model.

Our results demonstrate that MCMC methods for ODEs and SDEs provide practical algorithms for reconstruction transcription profiles whilst estimating some of the rate parameters involved. As the real population dynamics are naturally stochastic SDEs provide the superior theoretical model. However, the mean ODE approach can be useful as a vehicle for estimation when the data are not fully compatible with the SDE assumptions. Whilst they usually describe the same model in the mean, their difference lies in the specification of the variance. The SDE model provides a rigid description of the volatility process which is rigorously derived for the stochastic dynamics of the molecular processes. In theory, it is straightforward to allow for additive measurement error...
[see Heron et al. (2007) for estimation of SDEs with measurement error]. However, identification of an unknown measurement error variance is difficult and—to our knowledge—is not possible when the data are coarse and indirectly measured with unknown scaling factors. The algorithm for SDE estimation can be challenging to implement due to bridge sampling and is computationally expensive. Case 2 shows a problem that we have also encountered in Heron et al. (2007), namely if molecular populations are measured indirectly then the estimation of unknown scaling parameters can be difficult in practice. This may happen as a consequence of observing data that are too coarse, in the sense that too little information about the volatility process is revealed, or that are otherwise not directly compatible with the SDE assumption. However, drawbacks of the SDE approach are associated with the current quality, quantity and availability of the data. Case study 3 exemplifies that SDE estimation constitutes a very informative approach in calibrating all processes back to the molecular population levels as the scaling parameters can be identified. Under suitable assumptions the SDE model provides a theoretically well-founded modeling approach for describing the dynamics of molecular populations in a single cell. Estimation of SDEs is well studied and feasible and is highly informative when relatively frequent and clean (i.e. with little measurement error) single cell data are available on all model variables.

**Funding:** This research was funded by BBSRC and EPSRC (Interdisciplinary Programme for Cellular Regulation GR/S29256/01) and EU (BIOSIM Network Contract 005137). The Millar group’s work (K.E. and A.J.M.) was funded by BBSRC via award E015263 and CSBE (Centre for Integrative and Systems Biology funded by BBSRC and EPSRC). D.A.R. is funded by EPSRC Senior Research Fellowship EP/C544587/1 and M.K. by studentship, Department of Statistics, University of Warwick. C.V.H. was funded by Wellcome Trust Programme Grant (067252, to J.R.E.D. and M.R.H.W.) and now is recipient of The Prof. John Glover Memorial Postdoctoral Fellowship.

**Conflict of Interest:** none declared.

**REFERENCES**


Heron,E.A. et al. (2007) Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. *Bioinformatics*, **23**, 2589–2595.


