Bioinformatics Advance Access published October 30, 2008

Reconstruction of transcriptional dynamics from gene reporter data using differential equations

Bärbel Finkenstädt [1], *Elizabeth A. Heron [1,2],Michal Komorowski [1,2], Kieron Edwards [3],Sanyi Tang [2],†Claire V. Harper [4], Julian R. E. Davis [5], Michael R. H. White [4], Andrew J. Millar [3] and David A. Rand [2]

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

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: Promoter driven reporter genes, notably luciferase (luc) and green fluorescent protein (gfp), 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 and ordinary differential equations 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 stochastic or ordinary differential equations 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.

Supplementary Information: Supplementary information (SI) is provided with the submission.

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

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 data sets 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 data sets supports more or less complex modeling approaches we consider both stochastic and ordinary differential equations 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.

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

*to whom correspondence should be addressed

†Current Address: College of Mathematical and Information Science, Shaanxi Normal University, Xi'an, 710062 P.R. China

data with its limitations. Stochastic differential equations (SDEs) provide a good approximation of molecular population systems when one can assume that there is a macroscopic time scale for which (a) the event rates can be regarded as constant and (b) there are many events of each type. An example of formulating and fitting an autoregulatory feedback system with transcriptional delay as a system of SDEs can be found in (Heron et al. 2007). However, if the data are too sparsely sampled in time to reveal information about the volatility process, or if measurements are not realizations of the same continuous stochastic process in a cell, then the assumption of SDEs can be problematic in estimation. Simpler modeling approaches based on ODEs to represent the mean process with an additional stochastic error may provide a useful vehicle for estimation purposes at least in systems that have relatively regular and stable dynamics. The formulation of ODEs to model the dynamics of molecular population processes has become a widespread tool in systems biology (see, for example, systems studied in (Goldbeter 2002; Jensen et al. 2003; Locke et al. 2005a,b)), and early statistically less rigorous attempts in obtaining kinetic parameters from GFP reporter data can be found in (Ronen et al. 2002) and (Kalir and Alon 2004).

Here we consider the following dynamic model as the mechanistic backbone for the reconstruction of transcription profiles from reporter protein data

$$dM/dt = \tau(t) - \delta_M M(t), \quad dP/dt = \alpha M(t) - \delta_P P(t), \quad (1)$$

where M denotes the abundance of mRNA molecules and P denotes the abundance of the corresponding protein. The first equation describes the dynamics of mRNA molecules where transcription is given by a non-negative function $\tau(t)$. The second equation states that the protein is synthesized at a rate proportional to the abundance of mRNA. The mRNA and the protein are degraded (or leave their molecular compartment otherwise) at time scales with mean $1/\delta_m$ and $1/\delta_p$, respectively. The aim is to infer the transcription function $\tau(t)$ and possibly other rate constants of the system given time series data proportional to one or both variables of the system. Suppose that we measure M, P proportionally to their population size, $s_M M(t)$ for the mRNA and $s_P P(t)$ for the reporter protein. Re-parameterizing (1) gives a scaled model which is identical to (1) with scaled terms for α and τ (see SI). However, degradation rates are not affected by scaling. Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^T = \{M(t_i), P(t_i)\}_{i=1}^T$ denote experimental time series data observed at discrete time points. In order to obtain a likelihood function that incorporates the mechanistic rules in (1) we consider two approaches. One is the SDE approach where (1) is formulated as an appropriate system of stochastic differential equations. This approach is rigorously modeling the volatility of the stochastic dynamics of the kinetic processes provided that the assumptions of the SDE approximation itself are valid. It is very challenging to incorporate additional measurement error unless its variance is known or assumed. The second is the mean ODE approach where we assume that a solution path to (1) represents the mean of a stochastic process whilst the modeler makes assumptions about the probability distribution of the residual process. This approach is less exact than the SDE approach in modeling the volatility of the underlying stochastic interaction between molecules. On the other hand it naturally deals with measurement error and might also be useful for fitting to data sets which do not comply with the SDE assumption, for example, if data points are averages over replicates, come from different samples and/or represent a population of cells. We now introduce the two approaches and their likelihood derivation in more detail.

SDE approach: Here, M and P are random variables of molecular population sizes and the rates of increase and decrease in model (1) are event probabilities of birth and death processes at the individual molecular level. One can derive the following Itô SDEs (see SI)

$$\begin{aligned} dM &= \zeta_M(t,\theta)dt + \sigma_M(t,\theta)dW_M \\ dP &= \zeta_P(t,\theta)dt + \sigma_P(t,\theta)dW_P, \end{aligned} \quad (2)$$

where $\zeta_M(t,\theta) = \tau(t) - \delta_m M(t)$, $\zeta_P(t,\theta) = \alpha M(t) - \delta_P P(t)$, and $\sigma_M(t) = s_M^{1/2}(\tau(t) + \delta_m M(t))^{1/2}, \sigma_P(t) = s_P^{1/2}(\alpha M(t) + \delta_P P(t))^{1/2}$ are drift and volatility functions, respectively and W_M and W_P are independent Wiener processes [1]. Here and throughout the paper θ is used to denote a vector of model parameters. If M and P are indirect measurements of molecular populations in the sense that they are proportional to molecular abundance with factors s_M, s_P then these factors arise as additional parameters in the volatility functions and their estimation will be extremely useful allowing us to calibrate the model to the population level. Given data $\mathbf{Y}$ the likelihood function for the diffusion process is

$$L_{\text{SDE}}(\theta; \mathbf{Y}) = \prod_{i=1}^{T-1} f(\mathbf{y}_{i+1}|\mathbf{y}_i; \theta) \quad (3)$$

where $f(\mathbf{y}_{i+1}|\mathbf{y}_i;\theta)$ denotes the transition density of $\mathbf{y}_{i+1}$ given $\mathbf{y}_i$, that is the joint probability distribution of $M(t_{i+1})$ and $P(t_{i+1})$ given present values, under parameter vector θ. The exact transition density function for solutions of SDEs is rarely available in analytical form and usually approximations have to be considered. If the time-step $\Delta t_i = t_{i+1} - t_i$ is small then a good approximation is given by assuming that, conditional on past values,

(*) Increments $\mathbf{y}(t_{i+1}) - \mathbf{y}(t_i)$ are bivariate normal with mean vector $\zeta(t_i)\Delta t_i$ and variance matrix $\Sigma(t_i)\Delta t_i$ where $\zeta(t_i) = (\zeta_M(t_i), \zeta_P(t_i))$, $\Sigma(t_i) = \text{diag}(\sigma_M^2(t_i), \sigma_P^2(t_i))$ are the drift and volatility functions as defined above.

Thus, for sufficiently small sampling intervals Δt_i the likelihood function can be approximated by a product of the form

$$L_{\text{SDE}}(\theta; \mathbf{Y}) = \prod_{i=1}^{T-1} \Phi(\mathbf{y}(t_{i+1}) - \mathbf{y}(t_i); \zeta(t_i)\Delta t_i, \Sigma(t_i)\Delta t_i) \quad (4)$$

where $\Phi(x; \mu, \Sigma)$ denotes the bivariate normal density function with mean vector μ and variance matrix Σ. 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 (M_0, P_0). Then a natural probabilistic model is to assume that $\mathbf{Y}$ has a joint distribution with mean function $\mu(t;\theta)$ and a variance function $\sigma^2(t;\theta)$. The distribution function and

[1] The Wiener process, or Brownian motion, is a continuous-time stochastic process that has independent normally distributed increments.

variance are specified according to assumptions that the modeler makes about the residual process and measurement error. If the error process is assumed independent then the likelihood in the mean ODE approach is

$$L_{\text{ODE}}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{T} g(\mathbf{y}_i | \mu(t_i), \sigma^2(t_i), \theta), \tag{5}$$

where θ now incorporates initial conditions (M_0, P_0) and g is a suitably chosen probability distribution.

Inference: By Bayes' theorem the posterior distribution is

$$\pi(\theta|\mathbf{Y}) \propto L(\theta|\mathbf{Y})\pi(\theta), \tag{6}$$

where L is the likelihood function, derived for either the ODE or SDE approach, and $\pi(\theta)$ 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 twofold: Firstly, 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. Secondly, 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 (Elerian et al. 2001; Durham and Gallant 2002) 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 posterior $f(\theta, Y^*|Y)$ of the parameters θ and the latent variables Y^* given the data Y, using the fact that, by Bayes' theorem,

$$\pi(\theta, Y^*|Y) \propto L(Y^*, Y|\theta)\pi(\theta) \tag{7}$$

where $L(Y^*, Y|\theta)$ is the approximated augmented likelihood. This is achieved by sampling in turn from the full conditional densities of $\theta|Y^*, Y$ and $Y^*|\theta, Y$ (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 practice we have used (see (Heron et al. 2007)) a bridging method based on an independence sampler suggested by (Elerian et al. 2001)(see SI). 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 results we base our choice on Monte Carlo studies of simulated systems.

CASE STUDIES

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 SI 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 -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 SI) 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 seven hours. 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

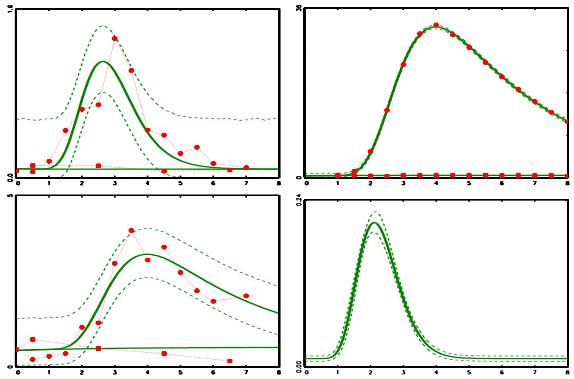

Fig. 1. This figure shows mean ODE fit for average data (data points given by big dots) of red light pulse experiment. *LUC* mRNA (top left), LUC activity in vitro (bottom left) and imaging the luminescence from LUC protein (top right) under two experimental conditions: with and without red light pulse. Solid lines give the mean ODE fit using mean posterior estimates for the parameters. The 95 % credible intervals (dashed lines) are shown for the control experiments. The reconstructed transcription profile $\tau(t)$ is shown in the bottom right panel (the area between dashed lines gives 95 % central values of the transcription profile for 10,000 iterations of Markov chain).

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

$$dP_v/dt = \alpha_{P_v} M(t) - \delta_P P_v(t), \tag{8}$$

which represents the protein dynamics measured by the in vitro LUC protein assays (see SI for full model statement). The two protein equations are identical except for differently scaled translation rates α_P and α_{P_v}.

Furthermore a constant c_P 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 $\tau(t)$ 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.* ,

$$\tau(t) = \alpha_M \left(\int_0^\infty d(u)L(t-u)du + \overline{\tau} \right), \qquad (9)$$

where $\overline{\tau}$ represents a baseline transcription. We take $d(u)$ to be a Gamma density with mean μ_Γ and standard deviation σ_Γ 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 is 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 $\sigma^2_M, \sigma^2_P, \sigma^2_{Pv}$. Applying (5) the likelihood of replicate $r = 1, 2, 3$ is

$$L^r(\boldsymbol{\theta}^r|\mathbf{Y}^r) = \prod_{i=1}^{T^R} g(\mathbf{y}_i^{r,R}|\mu(t_i), \theta^r) \prod_{j=1}^{T^C} g(\mathbf{y}_j^{r,C}|\mu(t_j), \theta^r), \qquad (10)$$

where $\mathbf{y}_i^{r,R}$ is the vector of observed data points $i = 1, ..., T^R$ for variables M, P, P_v for replicate r under the red light experiment, $\mathbf{y}_j^{r,C}$ 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 Fig. (1). The mean delay time between light induction and transcription is about 2h with almost all transcription happening between 0.8h and 3.2h after the pulse. Convergence of the Markov chains for parameters associated with the Gamma delay is relatively quick and precise. Chains for α_M and δ_M are correlated and convergence for these is slower. The half-life of *LUC* mRNA is estimated to be around 0.5 hours with some small variation between replicates. In contrast the chains for δ_L converged quickly due to the abundance and smoothness of the imaging data. Protein half-life was estimated to be around 2 to 2.5 hours. Although the control data do not seem very dynamic they are useful in inferring the base rates of transcription and translation. If the control series are omitted from the analysis these rates were estimated with considerably less precision and slower convergence due to correlations.

Table 1. Case 1: Posterior results for selected parameters.

Parameter	average	r1	r2	r3
δ_M	1.542 (0.019)	1.726 (0.044)	1.417 (0.121)	3.526 (0.315)
(half-life)	0.45 h	0.4 h	0.49 h	0.2 h
μ_Γ	2.008 (0.011)	2.101 (0.014)	1.902 (0.045)	2.362 (0.0289)
σ_Γ	0.631 (0.013)	0.692 (0.014)	0.686 (0.039)	0.723 (0.0217)
$\overline{\tau}$	0.012 (0.001)	0.014 (0.001)	0.014 (0.002)	0.013 (0.002)
δ_P	0.305 (0.0045)	0.286 (0.0040)	0.272 (0.010)	0.365 (0.0093)
(half-life)	2.27 h	2.42 h	2.5 h	1.9 h

Posterior means and standard deviations of selected estimated parameters (See SI for all parameters), where the red light pulse model was fitted to average data and to single replicate data sets 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 (per hour).

Case study 2: A Switch model for CCA1

The Circadian Clock Associated 1 (CCA1) gene in Arabidopsis thaliana has been identified as one of the core genes of the circadian clock (Wang and Tobin 1998). In this case study we show results for the reconstruction of an ON/OFF switching transcription profile from the following two experimental data sets:

(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 hours 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) was sampled at 1.5h 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 SI 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 3 variables where *LUC* mRNA and LUC protein dynamics are described by (1) and a further equation

$$dM_g/dt = \tau(t) - \delta_{M_g} M_g(t) \qquad (11)$$

is added for the native *CCA1* mRNA. We assume that observed variables are proportional to M_g and P populations with scaling factors s_{M_g} and s_P, while M is unobserved. To describe the oscillatory nature of the data we consider an ON/OFF switching function for the transcription $\tau(t) = \tau_{\text{on}}$ if transcription is active at time t, and $\tau(t) = \tau_{\text{off}}$ 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 $Sw = (s_1, ..., s_R)$ where $s_1 < s_2 < ... < s_R$ 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 here the number of switches and the initial state are known[3]. To set the phase of the clock both 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 $p_d\tau_{on}$ during the first 35 hours (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 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 as solution to an ODE from an initial condition M_0 to be estimated had no discernable impact on our

[2] For computational precision we amplified the mRNA concentrations by factor 10^5 and the protein concentrations by 10^4.

[3] The 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.

inference results here. In order to fit an SDE model to discrete data points for M_g and P 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 $\theta = (Sw, \tau_{on}, \tau_{off}, \delta_{M_g}, M_0, \delta_M, S_{M_g}, \alpha, \delta_P, S_P)$ denote the vector of unknown parameters and let M_g^* and P^* be the auxiliary data for M_g and P, respectively. Then according to (7) the posterior distribution for the unknown Θ, 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.* $\mathbf{y} = (M_g, P, M_g^*, P^*)$. More details of the SDE inference algorithm are provided in the SI.

Mean ODE approach: Here the likelihood is given by (5) where the unobserved variable M is reconstructed as a 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 $\sigma^2_{M_g}$ and σ^2_P. We have set $\tau_{\text{off}} = 0$ for the off-time as initial estimations showed that it was not different from zero[4]. 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 $\theta = (Sw, \tau_{\text{on}}, \tau_{\text{off}}, \delta_{M_g}, M_0, \delta_M, \alpha, \delta_P, \sigma_{M_g}, \sigma_P)$.

To ensure identifiability in both estimation approaches the prior distribution for *CCA1:LUC* mRNA degradation δ_M has to be informative. We hence used a Gamma distribution with mean 1.542 and standard deviation 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). Fig. (2) shows the transcription profiles 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) hours are at least twice as long as the active times (around 7 hours) 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 δ_{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, s_P and s_{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 s_{M_g} 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 Fig. (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 is a more aggregated and smoother time series than the mRNA series.

[4] We could not set $\tau_{\text{off}} = 0$ 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.

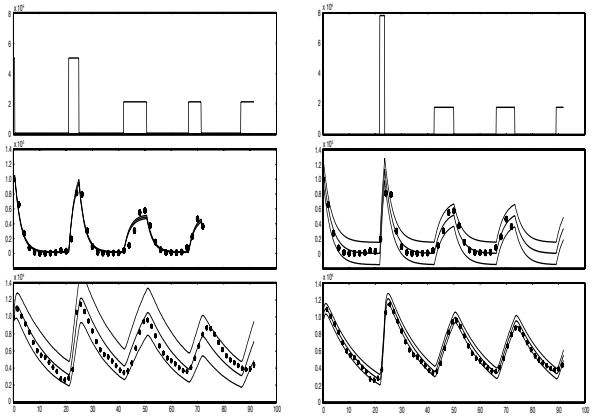

Fig. 2. Results of fitting SDEs (left) and ODEs (right) in case study 2. Top panel shows the mean reconstructed transcription profile $\tau(t)$ using the switch approximation. Middle panel shows results for M_g. Bottom panel gives results for P. Big dots are experimental data for M_g (middle panel) and P (bottom panel). The variation is shown as follows: For SDE approach (left): solid lines in middle and bottom panel give the 5 % , mean and 95 % values computed from 10,000 simulations of the SDE (using mean posterior parameter estimates). For ODE approach (right): Solid lines corresponds to the mean ODE fit (using mean posterior parameter estimates) plus/minus twice the mean posterior standard error.

Table 2. Case 2: Posterior results for selected parameters

	δ_{M_g}	δ_M	δ_P
SDE	0.426 (0.0043)	1.54 (0.019)	0.072 (0.0057)
ODE	0.313 (0.0273)	1.42 (0.101)	0.075 (0.0018)

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 SI.

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 5kb human prolactin gene promoter fragment linked to a destabilized EGFP reporter gene (hPRL-d2EGFP) (see SI for details of experiment). Images were taken 108 times in 15 minutes intervals giving a total of 27 hours of data for a single cell (see Figure (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 (δ_M, δ_P) and a strongly informative prior density is needed. Here we assume that they each have an independent Gamma distribution with mean 0.4 for δ_M and 0.5 for δ_P[5]. The prior variance was arbitrarily chosen to be small at 0.02 for both parameters. Since M is unobserved we can arbitrarily fix $s_M = 1$. Given the particular form of an experiment, where transcription is induced and afterwards comes back to its initial level, we have specified

[5] These rates were motivated by preliminary estimation using a small data set from other experiments. They are used here only to demonstrate the case as their estimates may change if more data were available.

$\tau(t)$ as follows

$$\tau(t) = \begin{cases} b_0 \exp(-\frac{(t-b_3)^2}{b_1}) + b_4 & t \leq b_3 \\ b_0 \exp(-\frac{(t-b_3)^2}{b_2}) + b_4 & t > b_3, \end{cases} \quad (12)$$

where the parameters b_i are to be estimated. Priors for parameters different than degradation rates were intended to be uninformative. Here we used exponential prior with means given in Table (3). The challenge for inference here is to integrate over a fully unobserved process M whilst sampling bridges 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 $\theta = (\delta_M, \delta_P, \alpha, s_P, b_0, b_1, b_2, b_3, b_4)$. The posterior distribution takes the form

$$\pi(\Theta, M^*, P^*|P) \propto L(M^*, P^*, P|\Theta)\pi(\Theta) \quad (13)$$

where we approximate $L(M^*, P^*, P|\Theta)$ with the likelihood (4) for the augmented data case, *i.e.* $\mathbf{y} = (M^*, P^*, P)$. 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 re-parameterizing the model (details of this and the algorithm are given in the SI). The algorithm was first tested on simulated data from the SDE model with chosen parameters (see Table (3)). Artificial data are simulated on a fine scale of 15/51 minutes and coarse data are extracted for P at 15 min intervals. The simulated and observed time series, and the reconstructed $\tau(t)$ are shown in Fig. (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 s_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.

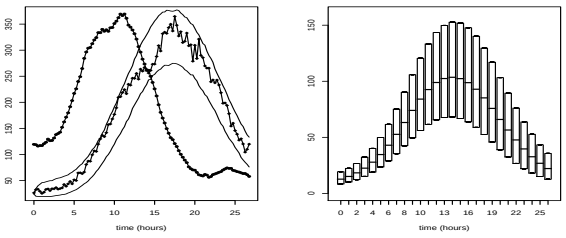

Fig. 3. Left: Time series of fluorescence intensity used in case study 3. Solid and dashed lines represent experimental and simulated data, respectively. The variation of the SDE fit to the real data is shown by the 5 % and 95 % values computed from 1,000 simulations of the SDE (using mean posterior parameter estimates). Right: Box-plot representing transcription profile in molecules per hour inferred from experimental data presented in the top figure. Each box represents 50% credibility interval and median of posterior distribution of the reconstructed transcription rate at particular time point.

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 re-parameterization of the model. We believe that the functional specifications for $\tau(t)$ 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 data sets 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.

Table 3. Case 3: Posterior inference results.

	value	prior	Simulation	Experiment
δ_M	0.44	Γ(0.44,0.02)	0.56 (0.36 - 0.92)	0.45(0.26 - 0.82)
δ_P	0.52	Γ(0.52,0.02)	0.59 (0.38 - 0.89)	0.71 (0.45 - 1.09)
α	20	Exp(100)	16.97 (6.54 - 78.98)	0.46 (0.14 - 1.51)
s_P	0.2	Exp(1)	0.17 (0.09 - 0.3)	2.11 (1.24 - 3.56)

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 μ and variance σ^2. Full list of all parameter estimates is provided in SI.

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 variance process of the mean ODE approach is not rigorously derived and can be specified by the modeler in an attempt to capture anything known about the residual process and measurement error. Estimation algorithms for the mean ODE approach are straightforward to implement although for higher dimensional or less stable systems more difficulties may occur. 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 practise. 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.

ACKNOWLEDGEMENTS

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 (KE and AJM) was funded by BBSRC via award E015263 and CSBE (Centre for Integrative and Systems Biology funded by BBSRC and EPSRC). DAR is funded by EPSRC Senior Research Fellowship EP/C544587/1 and MK by studentship, Dept of Statistics, University of Warwick. CVH was funded by Wellcome Trust Programme Grant (067252, to JRED and MRHW) and now is recipient of The Prof. John Glover Memorial Postdoctoral Fellowship.

Author Contributions

BF conducted the numerical estimations for case studies 1 and 2 (ODE part). EAH and MK did numerical estimations for case study 2 (SDE) and 3, respectively, under the supervision and guidance of BF and DAR. ST contributed to the algorithm development at the early stages of the paper. KE performed experiments for case study 1 and 2, under guidance of AJM. CVH performed the experiment for case study 3 under guidance of JRED and MRHW. BF wrote the paper with assistance from MK, EAH and DAR. DAR provided help on the mathematical modeling and initiated the collaboration between the theoretical and experimental groups.

REFERENCES

S. O. Aase and P. Ruoff. Semi-algebraic optimization of temperature compensation in a general switch-type negative feedback model of circadian clocks. *Journal of Mathematical Biology*, 56(3):279–292, 2008.

G. B. Durham and A. R. Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–316, 2002.

O. Elerian, S. Chib, and N. Shephard. Likelihood inference for discretely observed non-linear diffusions. *Econometrica*, 69:959–993, 2001.

D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference, 2nd ed.* Chapman & Hall/CRC, Boca Raton, London, New York, 2006.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188 (1-3):404–425, 1992.

A. Goldbeter. Computational approaches to cellular rhythms. *Nature*, 420(6912):238–245, 2002.

S. Hengl, C. Kreutz, J. Timmer, and T. Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007.

E. A. Heron, B. Finkenstädt, and D. A. Rand. Bayesian inference for dynamic transcriptional regulation; the hes1 system as a case study. *Bioinformatics*, 23(19): 2589–2595, 2007.

M. H. Jensen, K. Sneppen, and G. Tiana. Sustained oscillations and time delays in gene expression of protein $Hes1$. *Febs Letters*, 541(1-3):176–177, 2003.

S. Kalir and U. Alon. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. *Cell*, 117:713–720, 2004.

P. E. Kloeden and E Platen. *Numerical solution of stochastic differential equations*. Springer-Verlag, 3rd Ed., Berlin; New York, 1999.

J. C. W. Locke, A. J. Millar, and M. S. Turner. Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *J Theor Biol*, 234:383–393, 2005a.

J. C. W. Locke, M. M. Southern, L. Kozma-Bognar, V. Hibberd, P. E. Brown, M. S. Turner, and A. J. Millar. Extension of a genetic network model by iterative experimentation and mathematical analysis. *Mol Syst Biol*, 1:E1–E9, 2005b.

A. J. Millar and S. A. Kay. Integration of circadian and phototransduction pathways in the network controlling cab gene transcription in arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 93:15491–96, 1996.

A. J. Millar, I. Carré, C. Strayer, N. Chua, and S. Kay. Circadian clock mutants in arabidopsis identified by luciferase imaging. *Science*, 267:1161–1163, 1995.

D. E. Nelson, A. E. C. Ihekwaba, M. Elliott, J. R. Johnson, C. A. Gibney, B. E. Foreman, G. Nelson, V. See, C. A. Horton, D. G. Spiller, S. W. Edwards, H. P. McDowell, J. F. Unitt, E. Sullivan, R. Grimley, N. Benson, D. Broomhead, D. B. Kell, and M. R. H. White. Oscillations in NF-kappa B signaling control the dynamics of gene expression. *Science*, 306(5696):704–708, 2004.

M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulatory network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences*, 99(16):10555–60, 2002.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

Z. Wang and E. M. Tobin. Constitutive expression of the circadian clock associated 1 (cca1) gene disrupts circadian rhythms and suppresses its own expression. *Cell*, 93 (7):1207–1217, 1998.

E. Yakir, D. Hilman, M. Hassidim, and R. Green. Circadian clock associated1 transcript stability and the entrainment of the circadian clock in arabidopsis. *Plant Physiology*, 145:925–932, 2007.