Bioinformatics Advance Access published July 28, 2007

Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study.

Elizabeth A. Heron [a,b], Bärbel Finkenstädt [b*], David A. Rand [a,c]

[a]Warwick Systems Biology Centre,[b]Department of Statistics, [c]Department of Mathematics,
University of Warwick, Coventry CV4 7AL, UK

Associate Editor: Dr. Limsoon Wong

ABSTRACT

Motivation: In this study we address the problem of estimating the parameters of regulatory networks and provide the first application of Markov chain Monte Carlo (MCMC) methods to experimental data. As a case study we consider a stochastic model of the Hes1 system expressed in terms of stochastic differential equations (SDEs) to which rigorous likelihood methods of inference can be applied. When fitting continuous-time stochastic models to discretely observed time series the lengths of the sampling intervals are important, and much of our study addresses the problem when the data are sparse.

Results: We estimate the parameters of an autoregulatory network providing results both for simulated and real experimental data from the Hes1 system. We develop an estimation algorithm using Markov chain Monte Carlo techniques which are flexible enough to allow for the imputation of latent data on a finer time scale and the presence of prior information about parameters which may be informed from other experiments as well as additional measurement error.

Availability: Supplementary information is submitted with the paper.

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

1 INTRODUCTION

While there are now a large number of important models for regulatory and signalling systems, few of these have been systematically fitted to data. Indeed, this is a very challenging problem because, while even the simplest realistic networks involve a large number of parameters, the time series data available are typically noisy, indirect and of limited time resolution. In addition, the dynamical systems arising from regulatory and signalling networks are strongly nonlinear, have high-dimensional state spaces and are intrinsically stochastic. On the other hand the ability to infer the parameters of such networks will become increasingly important if the study of biological systems is to become more quantitative and predictive. It is this problem that we address here by developing a method based on Markov chain Monte Carlo (MCMC) techniques to estimate the parameters of simple networks modelled by stochastic ordinary differential equations and by applying this to the Hes1 network.

The literature now contains a broad array of explicitly molecular models for regulatory and signalling systems that are described by differential equations (both with and without delay). A typical Ansatz is a compartmental population model based on ordinary differential equations (ODEs) where the different molecular species represent the state variables. In order to generate oscillations these often involve negative feedback, sometimes combined with delays due to the time taken for transcription, translation, post-translational modification, transportation etc. This, for example, is a common feature of models of circadian clocks (Goldbeter 2002), the p53 system (Haupt et al. 1997), the NF-κB system (Hoffmann et al. 2002) and the Hes1 system (Hirata et al. 2002).

However, as is now well understood, 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 such systems are intrinsically stochastic (McAdams and Arkin 1997; Koern et al. 2005). Moreover, the data for such systems reflect this stochasticity and are affected by measurement noise and other uncertainties. Therefore, for an approach like ours which is based on exploiting the probabilistic structure, it is necessary to formulate stochastic models for these systems. To do this 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. However, such models are too detailed for there to be any hope of fitting to current data with its limitations. In between ODEs and these models are stochastic differential equations (SDEs) that have the advantage of being defined by functional relationships like ODEs but are also able to incorporate the effects of stochasticity. SDEs provide a good approximation of the full stochastic systems when there is a macroscopic time scale with the property that during such a period (a) the event rates can be regarded as constant and (b) there are many events of each type. Our methods to estimate parameter values use such models.

Biological data often suffer from having too coarse a time resolution to allow parameter estimation in a straightforward way. The methods applied in this study make use of strategies developed for nonlinear stochastic differential equations, which have received much attention in the econometric literature in particular in the context of stochastic volatility models in finance. An important approach firstly adopted by Pedersen (1995) is to augment the observed data by introducing a number of latent data points in-between the measurements. This idea has been further pursued by (Kim et al. 1998; Eraker 2001; Elerian et al. 2001; Durham and Gallant 2002) using simulated maximum likelihood estimation and/or Markov chain Monte Carlo (MCMC) algorithms to sample the posterior distributions of the parameters and the latent data in a Bayesian framework. The first published use of such an approach for the analysis of stochastic kinetic biochemical network models was in (Golightly and Wilkinson 2005, 2006). They considered a stochastic model of the detailed kinetic molecular interactions in a particular theoretical negative feedback loop, determined the correponding SDE and investigate the effectiveness of their MCMC algorithm with synthetic data.

[*]to whom correspondence should be addressed

It is the aim of this study to develop such a statistical framework for stochastic differential equations describing a delayed autoregulatory feedback loop and provide the first application of these methods to experimental data that suffer from some real shortcomings. As a case study we consider the Hes1 system using the experimental time series data provided by Hirata et al. (2002). We demonstrate the power and the usefulness of the statistical methodology by applying it to simulated data from current Hes1 models and then to experimental data. The experimental data we use, from Hirata et al. (2002), are extremely sparse, indirect and possibly also subject to measurement error and we have extended our methodology as far as possible to cope with these problems.

Hes1 oscillations are thought to provide a clock for the segmentation of the somites during embryogenesis. A series of experiments by Hirata et al. (2002) demonstrates that the Hes1 system, which oscillates at a period of about 2 hours, could essentially be as simple as a single feedback loop in which the protein product of the *hes1* gene acts so as to repress *hes1* transcription. Time series observations, albeit sparse, on the relative concentrations are available for the mRNA and protein concentration (Hirata et al. (2002)) (Figure 1).

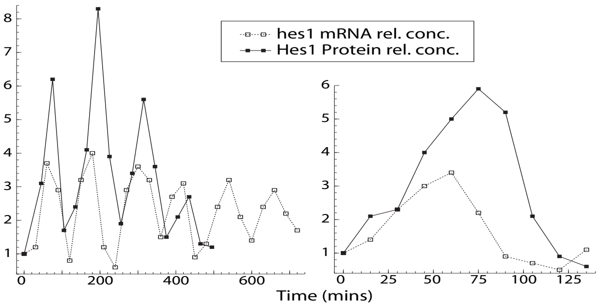

Fig. 1. The left and right panel give two time courses of relative concentrations of Hes1 mRNA and protein obtained in Hirata et al. (2002). The observed data are given by the discrete points in the plots which are connected only for illustration. In one experiment (left panel), 17 discrete data points are available that describe the contemporaneous time course of the mRNA and protein at 30 min long time intervals - except the first protein measurement was taken 45 min after an initial measurement at time 0. Protein and mRNA are not measured at the same time but are 15 minutes apart. A further time course for both variables (right panel) with 10 data points measured at 15 min interval length was also obtained by Hirata et al. (2002). Both data sets are used in the estimations.

Hirata et al. (2002) concluded that there must be an additional unobserved Hes1 interaction factor because the relevant systems of two ODEs cannot generate sustained oscillations. Subsequent mathematical analyses by Monk (2003), Lewis (2003), Jensen et al. (2003), Bernard et al. (2006), and Barrio et al. (2006) showed that this extra factor was not necessary because a model with just a single feedback loop with the sort of biological process delays mentioned above can readily and robustly generate sustained oscillations. For the Hes1 system the transcriptional delays are thought to be in the range of 10 to 20 minutes (Monk 2003) whilst the translational delays are about 1-3 minutes (Monk 2003). Recently, Bernard et al. (2006) compared two modelling approaches for the Hes1 system, a simple feedback loop with discrete delay time and a 2-dimensional system including two proteins, Hes1 and Gro/TLE1 incorporating one delay. They conclude in favour of the latter because the additional nonlinear complexity prevents overshooting of the mRNA level and allows for some more realistic fine-tuning of the oscillations.

2 SYSTEM AND METHODS

2.1 Deterministic models

Consider the following two compartmental system of differential equations for the feedback loop

$$\frac{dM}{dt} = \frac{v_1}{1+(D[P(t)]/k_1)^n} - v_2 M(t), \quad (1)$$

$$\frac{dP}{dt} = v_3 M(t) - v_4 P(t). \quad (2)$$

Here M denotes *hes1* mRNA and P denotes Hes1 protein. The first equation describes the temporal variation of the number of mRNA molecules which are regulated by some delayed nuclear protein $D[P(t)]$ through a Hill function with parameters v_1, n and k_1. Here, D denotes some functional operator acting on past and present values of P. 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 possibly leave their molecular compartment otherwise) at time scales with mean $1/v_2$ and $1/v_4$, respectively. Hirata et al. (2002) estimate the half-lives of *hes1* mRNA and Hes1 protein to be 24.1 ± 1.7 min, 22.3 ± 3.1 min, respectively. We will incorporate this information into the analysis through prior parameter distributions.

The two cases we will be interested in are the case of a single discrete delay where $D[P(t)] = P(t-\tau)$ (Monk 2003; Lewis 2003; Bernard et al. 2006) and the general case of a distributed delay (Monk 2003) where

$$D[P(t)] = \int_0^\infty P(t-s)g(s)\,ds. \quad (3)$$

These are used to model the delays caused by the various molecular processes (translation, transcription, post-transcriptional modification, transport etc) that take place before the protein can act as a transcription factor. A natural choice for the distribution $g(s)$ is the gamma density with mean μ_g and variance σ_g^2. This is exact if all biological processes involved have independent identical exponentially distributed waiting times. However, our practical experience is that even if this assumption does not hold, the gamma density arises as a good pragmatic choice as it sufficiently general to approximate a range of densities and produces realistic oscillations in simulated clock time series.

2.2 Stochastic differential equations

Consider a birth-death process $Y(t)$ with birth rate $\beta = \beta(t)$ and death rate $\delta = \delta(t)$. Suppose that for a short time ΔT the rates hardly change and many birth and death events take place. Then one can show (see supplementary Section 1) that

$$\Delta Y(t) = Y(t+\Delta T) - Y(t) \sim N((\beta-\delta)\Delta T, (\beta+\delta)\Delta T)$$

i.e. $\Delta Y(t)$ is normally distributed with mean $(\beta - \delta)\Delta T$ and variance $(\beta + \delta)\Delta T$. We therefore model the Hes1 mRNA dynamics using a SDE of the form $dM = (\beta_M - \delta_M)\,dt + \sqrt{(\beta_M + \delta_M)}\,dW_M$ where the birth and death terms β_M and δ_M are given by the deterministic equation (1) above. Here dW_M represents the increments of a one-dimensional Brownian motion. A similar equation is used for the protein dynamics.

We modify this equation to take account of the fact that (as is usually the case for such systems) the time series data do not give absolute values of mRNA or protein numbers but only measure these up to proportionality. Thus, it is assumed that the observed variables are proportional to the original variables: $m(t) = s_M M(t)$ for the mRNA, and $p(t) = s_P P(t)$ for the protein. We therefore reformulate the equations in these observed variables, but for convenience we reuse the variables M and P which are henceforth used to denote the scaled variables. In this way we obtain a new system of equations

$$\begin{aligned} dM &= \mu_M\,dt + \sigma_M\,dW_M, \\ dP &= \mu_P\,dt + \sigma_P\,dW_P, \end{aligned} \tag{4}$$

where dW_M and dW_P are independent increments of one-dimensional Brownian motions. The detailed functional form of equation (4) is given in the supplementary information. At time t, μ_P and σ_P depend upon $M(t)$, $P(t)$ and the parameters and, because of the discrete or distributed delay, μ_M and σ_M also depend upon $P(s)$ for all or some $s \leq t$.

These equations (4) have the two new scaling parameters s_M and s_P as well as the scaled parameters $\tilde{v}_1 = s_M v_1$, $\tilde{k}_1 = s_P k_1$ and $\tilde{v}_3 = (s_P/s_M)v_3$, plus the original parameters $\tilde{v}_2 = v_2, \tilde{n} = n, \tilde{\tau} = \tau$ and $\tilde{v}_4 = v_4$. Let the vector $\theta = (\theta_M, \theta_P)$ summarize all parameters with $\theta_M = (n, \tilde{v}_1, v_2, \tilde{k}_1, \tau, s_M)$, the parameters for the M equation in the discrete delay case, and $\theta_P = (\tilde{v}_3, v_4, s_P)$, the parameters for the P equation. For the distributed delay model τ is replaced by μ_g and σ_g.

3 METHODS OF INFERENCE AND ALGORITHM

Suppose that we observe a discretely sampled multivariate time series $Y = \{(M(t_i), P(t_i)), i = 1, ..., T\}$ from the process. For simplicity and to help the exposition we will assume that the times at which M and P are observed are identical, though this does not have to be the case. The aim here is to estimate the parameters θ given the data Y through the posterior conditional distribution $f(\theta|Y)$. We firstly discuss this ignoring measurement error and then explain how measurement error is incorporated into our algorithm.

In order to perform likelihood based inference we need the transition density, that is the probability distribution of $M(t)$ and $P(t)$ given past values. The exact likelihood for solutions of SDEs is only 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,

(*) The increments $M(t_{i+1}) - M(t_i)$ and $P(t_{i+1}) - P(t_i)$ are normally distributed with mean and variance μ, σ^2 given respectively by $\mu_{M,i}, \sigma^2_{M,i}$ and by $\mu_{P,i}, \sigma^2_{P,i}$

where $\mu_{P,i}$ and $\sigma^2_{P,i}$ are functions of $M(t_i)$, $P(t_i)$, and θ_P and $\mu_{M,i}$ and $\sigma^2_{M,i}$ are functions of $M(t_i)$, $P(\leq t_i) = (P(t_i), P(t_{i-1}), \ldots P(t_{i-s}))$ and θ_M as in model (4). The latter dependence upon $P(\leq t_i)$ is due to the delay in the equation. In fact, the continuous distributed delay formally involves infinitely many previous times but one can safely truncate this.

Suppose that we have a given data set $Y = (M, P)$ as above and the sampling intervals Δt_i are sufficiently small that (*) holds. Let $L_{\text{SDE}}(\theta; Y)$ denote the likelihood of Y arising from a trajectory of the SDE when the parameters are $\theta = (\theta_M, \theta_P)$. This can be approximated by a product of the form

$$\begin{aligned} L_{\text{SDE}}(\theta; Y) \;=\; & \prod_{i=1}^{T} \Phi(M(t_{i+1}) - M(t_i); \mu_{M,i}, \sigma^2_{M,i}) \\ & \prod_{i=1}^{T} \Phi(P(t_{i+1}) - P(t_i); \mu_{P,i}, \sigma^2_{P,i}), \end{aligned} \tag{5}$$

where Φ is the Normal distribution i.e. $\Phi(x; \mu, \sigma^2) = (1/\sqrt{2\pi\sigma^2})\exp(-(x - \mu)/2\sigma^2)$. Justifications for such an approximation are given in Kloeden and Platen (1999). By Bayes' theorem the posterior distribution is given by

$$f(\theta|Y) \propto L_{\text{SDE}}(\theta; Y)\pi(\theta_M)\pi(\theta_P),$$

where $\pi(\theta_M)$ and $\pi(\theta_P)$ are prior distributions on the parameters of the M and P equations i.e. distributions that reflect our prior knowledge about the parameters.

3.1 Sparse data

Unfortunately, the experimental data of Hirata et al. (2002) are sparse i.e. Δt_i is not small (Figure 1). There exist various approaches in the literature that attempt to deal with such a situation (see, for example, Durham and Gallant (2002) for an overview). One simple idea leading to simulation based likelihood inference methods and/or MCMC is to augment the data by introducing a finer set of times $\tau_{i,j}$ so that each interval $[t_i, t_{i+1}]$ is partitioned into $F + 1$ subintervals $[t_i = \tau_{i,0}, \tau_{i,1}, \ldots, \tau_{i,F+1} = t_{i+1}]$. We will consider imputed data at the new times $\tau_{i,j}$ which we will denote by $M^*(\tau_{i,j})$ and $P^*(\tau_{i,j})$, $j = 1, \ldots, F$. These will be treated as latent (unobserved) variables and the imputed data between t_i and t_{i+1} is called a *bridge*. Let M^*_i and P^*_i denote the bridge vectors $(M^*(\tau_{i,1}), \ldots, M^*(\tau_{i,F}))$ and $(P^*(\tau_{i,1}), \ldots, P^*(\tau_{i,F}))$, then $Y^*_i = (M^*_i, P^*_i)$ and $Y^* = (Y^*_1, \ldots, Y^*_{T-1})$.

The new times are chosen so that the constant-rate first-order approximation (*) can be safely assumed to be accurate on each subinterval $[\tau_{i,j}, \tau_{i,j+1}]$. We can then use equation (5) to obtain an augmented approximate likelihood. In order to estimate a transition probability like $\pi(M(\tau_{i,j+1})|M(\tau_{i,j}), P(\leq \tau_{i,j}); \theta)$ we can use (5) to write them in terms of the variables in Y^* and then integrate these auxiliary variables out. Monte Carlo methods, where the unobserved processes on the subintervals are simulated as auxiliary random variables, provide a feasible way to perform this integration.

Thus, to provide an estimate of θ from sparsely sampled data, we use MCMC to sample from the joint posterior $f(\theta, Y^*|Y)$ of the parameters θ and the auxiliary variables Y^* given the data Y, using the fact that, by Bayes' theorem,

$$f(\theta, Y^*|Y) \propto L_{\text{SDE}}(Y^*, Y|\theta)\pi(\theta) \tag{6}$$

where, as before, $\pi(\theta)$ denotes the prior distribution on θ and $L_{\text{SDE}}(Y^*, Y|\theta)$ 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)). The general structure of the algorithm that we employ is thus as follows:

1. Initialise Y^* and the parameters θ.
2. Sample M_i^* from $M_i^*|M(t_i), M(t_{i+1}), \theta_M$ and sample P_i^* from $P_i^*|P(t_i), P(t_{i+1}), \theta_P$ for $i = 1, 2, \ldots, T-1$. The two samples constitute a full set of imputed data Y^*.
3. Sample θ from $\theta|Y, Y^*$, i.e. use the fully augmented data to update the parameter vector.
4. Repeat steps 2 and 3 until the required sample is obtained after the chain has converged.

Step 3, sampling from the conditional distribution, is described in more detail in the supplementary information. Updating the parameter vector is quite straightforward as for a given fully augmented time path the above constant rate approximation for L_{SDE} is valid and the inference problem is the same as for a finely sampled time path.

To sample M_i^* and P_i^* in step 2 the bridging methodology suggested by Elerian et al. (2001) has proved very satisfactory. A brief introduction to their method is available in the supplementary information. However, it should be noted that there exist various other available methods for bridging (see Durham and Gallant (2002) for a survey) that may also be used for this kind of problem.

3.2 Measurement error

We also want to allow for the fact that we expect that the data is subject to a significant amount of measurement error. We are not going to try and estimate the error but to take such error into account in the algorithm to ensure that our estimates are robust to its addition. To do this we will need to allow the data points $Y(t_i)$ to be moved by the algorithm. Therefore to explain this we temporarily make a slight change to the notation and let $X(t_i) = (M_{\mathrm{ob}}(t_i), P_{\mathrm{ob}}(t_i))$ stand for the observed data. We assume the simplest model for measurement error, namely that the difference between the true levels $M(t_i)$ and $P(t_i)$ and the observed levels $M_{\mathrm{ob}}(t_i)$ and $P_{\mathrm{ob}}(t_i)$ are normally distributed with mean zero and respective variances $\sigma^2_{e,M}$ and $\sigma^2_{e,P}$. To incorporate this in our Markov chain we replace step 2 above by 2':

2'. Use a Meteropolis-Hastings sampler where, firstly, changes to $Y(t_i)$ are proposed from the above normal distributions for the measurement errors and, secondly, changes to the adjoining bridges $Y^*(t_i)$ (if $i \neq 1$) and $Y^*(t_i)$ (if $i \neq T-1$) are proposed conditional on the new proposed value of $Y(t_i)$. The overall change is then accepted or rejected according to the appropriate Meteropolis-Hastings criterion.

Further details about the algorithm and its practical implementation can be found in the supplementary information.

4 ESTIMATION RESULTS

4.1 Simulated data from Hes1 system

When developing statistical techniques like ours it is crucial to test the algorithm on a range of simulated data in order to develop a good understanding of its effectiveness and reliability and the dependence of the results upon the type of data used.

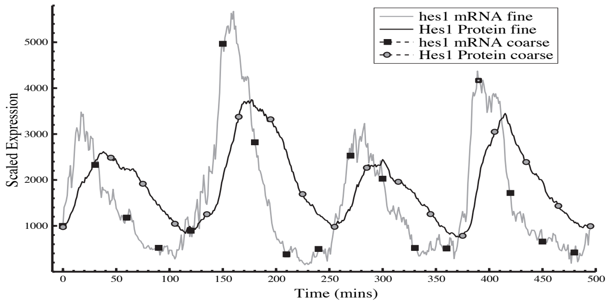

Fig. 2. Scaled Hes1 mRNA and protein data simulated using the parameters given in Monk (2003)($n = 5, v_1 = 1, v_2 = 0.03, k_1 = 100, \tau = 18, v_3 = 1, v_4 = 0.03$, scaling: $\tilde{M} = M(1000/M_0) = M(290.3), \tilde{P} = P(1000/P_0) = P(10.399)$). The data that are simulated and sampled at gridsize 1 minute are referred to as fine data. For the coarse data the same simulated data set is used but sampled discretely at the same time points as one of the data sets of Hirata et al. (2002) (left panel of Figure 1).

Figure 2 shows some simulated scaled *hes1* mRNA and protein data from the delay SDE model in equation (4) with fixed delay $D[P(t)] = P(t-\tau)$ and with the parameter values taken from Monk (2003). A first order Euler approximation on a gridsize of 1 minute is used. The length of the simulated data is 495 minutes corresponding to four 2-hourly oscillations. Simulated abundance was taken relative to the initial conditions and also multiplied by 1000 (this serves to prevent the sampled bridges from having negative components).

Similar simulated data was also generated (see supplementary Figure 1) for the distributed delay model (4) where the waiting time s has a probability distribution $g(s)$ and the amount of active transcription factor is as specified in (3). In particular we take $g(s)$ to be a gamma density. The parameter values were set to values obtained from fitting the fixed delay model to the experimental data of Hirata et al. (2002) where the posterior median of the fixed delay was estimated to be about 25 minutes (see Table 2). This was replaced by a distributed delay using a gamma distribution with mean 25 and standard deviation of 1 which is nearly a symmetric distribution with almost all probability mass between 22 and 28 minutes thus occupying about one fifth of the time of a single oscillation. We used the same initial conditions for M and P as in the previous section giving similar scaling coefficients.

4.1.1 Inference for finely sampled data Inference results using MCMC applied to the finely sampled data (Figure 2) from the discrete delay model are given in Table 1 and Figures 3 and 4. Here we omitted step 2 of the algorithm. All prior distributions were chosen to be uniform and each element of the parameter vector θ was updated using a random walk Metropolis algorithm (see supplementary information). The results show that informative posterior distributions can be readily obtained for all parameters in θ. This also includes both scaling parameters. Since these relate the observed data to the population levels of the molecular species, knowledge of their values allows us to estimate molecular population sizes. Similarly we can also retrieve all parameters well if we use the simulated fine data from the distributed delay model and fit to this model in a similar way (see supplementary Table 1 and supplementary Figure 2).

4.1.2 Inference for sparsely sampled data We now study what limitations hold for sparse data. Figure 2 also shows an example

Table 1. Results of fitting the discrete delay model to simulated data.

Parameter	set value	Fine Data	Coarse Data
n	5	6.26 (4.61 - 8.98)	5.57 (4.73 - 6.66)
$\tilde{v}_1$	290.3	237.7 (165.0 - 336.9)	324.9 (275.4 - 385.3)
v_2	0.03	0.0371 (0.0273 - 0.0464)	0.0424 (0.0386 - 0.0470)
$\log(\tilde{k}_1)$	6.95	7.096 (6.914 - 7.226)	6.964 (6.891 - 7.038)
τ	18	18 (16 - 21)	17 (15 - 18)
$\tilde{v}_3$	0.0358	0.0369 (0.0339 - 0.0397)	0.0348 (0.0314 - 0.0382)
v_4	0.03	0.0312 (0.0287 - 0.0336)	0.0298 (0.0271 - 0.0327)
S_M	290.3	224.3 (172.2 - 314.2)	-
S_P	10.4	9.41 (8.14 - 10.89)	-

Column 2 gives the true parameter values used to generate the simulated data. Columns 3 and 4 give sample median and 95% credible intervals for posteriors estimated from simulated data from the discrete delay model (shown in figure 2). Fine data are simulated and sampled at 1 minute intervals with a total of 495 time points (4 cycles) for each of the M and P time series. Coarse data are sampled at intervals corresponding to the data available in Hirata et al. (2002) (left panel of Figure 1).

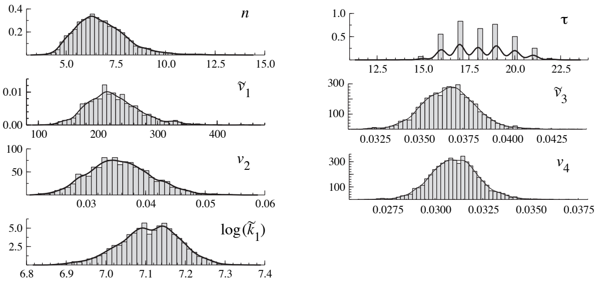

Fig. 3. Density estimates (posterior distributions) for all parameters for simulated M and P data on a fine time grid (gridsize of 1 minute). Simulation parameter values: $n = 5, \tilde{v}_1 = 290.3, v_2 = 0.03, \log(\tilde{k}_1) = 6.947, \tau = 18, \tilde{v}_3 = 0.0358, v_4 = 0.03$, values for the delay parameter are discretized into units of one minute.

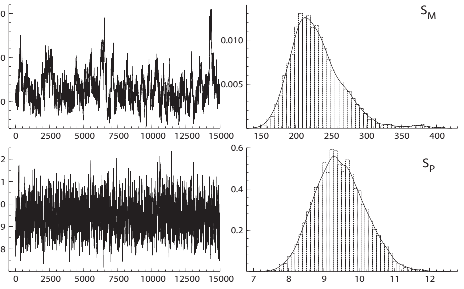

Fig. 4. Markov chains and density estimates (posterior distributions) for the scaling parameters s_M and s_P (results for the same data as used in Figure 3). Simulation parameter values $s_M = 290.3, s_P = 10.399$.

of a sparse discrete time series for the discrete delay model. This is obtained by sampling the fine data on a sparse grid as for the experimental data of Hirata et al. (2002). One can see intuitively that the detailed structure of the volatility of the stochastic process is almost entirely lost with such sparse sampling. Since the scaling factors tune the amplitude of the volatility, it is not reasonable to expect to be able to reconstruct the scaling parameters if the sampling is too coarse. Table 1 gives results for inference on the parameters θ using only the simulated coarsened data as shown in Figure 2. Plots of the estimated posteriors are provided in supplementary Figure 3. In this case the MCMC algorithm was extended to sample auxiliary data on a gridsize of 1 minute using the bridge building methodology based on an independence sampler as suggested by Elerian et al. (2001). The estimated posterior distributions are now substantially wider which is not surprising as we attempt parameter estimation using only 17 data points per variable. This is less than 5% of the data used in the previous estimation.

Nevertheless the Markov chains do converge to posterior densities that are consistent with the true scaled parameters while the scaling parameters drift within allowable bounded regions. Moreover, if the scaling parameters are held fixed within these regions the posterior distributions of the other parameters converge to values that only depend very weakly upon the values at which the scaling parameters are held fixed. Thus it seems that enough information is contained in the mean behaviour of the series to make parameter estimation feasible apart from the scaling coefficients.

Of course one cannot expect the true values of the parameters to necessarily correspond to the sample means of the calculated posterior distributions. This is because the data is a single sparse realisation of the SDE which may well have a non-optimal likelihood for the true parameters. Figure 5 shows the Markov chains for the parameters $\tilde{\nu}_3$ and ν_4. It is clear that these are highly correlated. This is a common occurrence for such systems whenever one parameter can play off against another. To understand this better consider the likelihood $L(\theta|Y^*)$ for a given Y^*. Let θ^* be the global maximum of this. We consider varying just two of the parameters, θ_i and θ_j, while keeping all the others fixed at the values θ^*_k of the maximun θ^*. Then the curves of constant likelihood in the θ_i, θ_j-plane close to the maximum are approximately ellipses. For the parameters $\tilde{\nu}_3$ and ν_4 they are long thin ellipses with the major axis close to the line $\tilde{\nu}_3 = \alpha\nu_4$ for $\alpha > 0$ of order one. The minor axis is very short. As the MCMC algorithm is iterated the parameter values run along the major axes of these ellipses giving the large excursions seen in the Markov chain. If however we plot the projection of these planes on to a line close to the minor axes (in this case by plotting $\tilde{\nu}_3/\nu_4$) we remove this drift and see more clearly that the chains have converged. These correlations can be removed by using an independence sampler that takes such correlations into account such as a version for parameters of the independence sampler detailed in Elerian et al. (2001) that we use for bridges.

Similar results hold for the fitting of the distributed delay model to coarse simulated data (see supplementary Table 1 and supplementary Figure 4 for detailed results and plots of estimated posterior distributions). The key difference is that the chain for σ_g now does not converge (see supplementary Figure 4). Nevertheless, even though σ_g and the scaling parameters do not appear to have converged the chains for all the other parameters converge to posterior densities that are consistent with the true parameters.

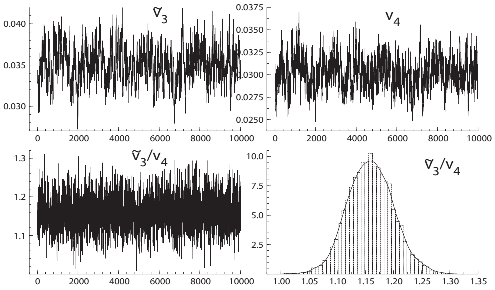

Fig. 5. This figure shows the correlation between the Markov chains for $\tilde{\nu}_3$ and ν_4. Plotting the chain and posterior density for $\tilde{\nu}_3/\nu_4$ strongly suggests that the chains have converged.

Table 2. Results for Hirata experimental data.

	discrete	with error	distributed
n	8.92 (5.67 - 15.66)	9.26 (5.86 - 20.13)	10.27 (5.83 - 34.22)
$\tilde{v}_1$	122.2 (107 - 139)	122.1 (108 - 139)	121.0 (107 - 139)
v_2	0.036 (0.032 - 0.040)	0.037 (0.033 - 0.041)	0.036 (0.032 - 0.040)
$\log(\tilde{k}_1)$	8.25 (8.16 - 8.34)	8.27 (8.19 - 8.35)	8.26 (8.16 - 8.35)
$\tilde{v}_3$	0.0875 (0.078 - 0.101)	0.087 (0.076 - 0.098)	0.0873 (0.077 - 0.010)
v_4	0.0563 (0.050 - 0.064)	0.056 (0.050 - 0.063)	0.057 (0.051 - 0.064)
τ	25 (22 - 28)	25 (22 - 28)	-
μ	-	-	26.31 (23.68 - 28.88)
σ	-	-	2.74 (0.925 - 5.330)

Sample median together with 95% credible intervals for each of the parameters of the discrete delay model and the distributed delay model using the data available in Hirata et al. (2002). The central colum gives results for the discrete delay model accommodating for measurement error in the bridges, allowing for Normal(0,500) error for both M and P.

4.1.3 Results for experimental data We apply the MCMC algorithm exactly as for the simulated coarsened data above, using the experimental data of Hirata et al. (2002) shown in Figure 1. Care has to be taken with respect to some unequal interval sizes and to the fact that M and P are not observed at the same time points. Originally we started by using only the data given in the left panel of Figure 1 but we found that including the data of the single additional oscillation shown in the right panel of Figure 1 improves the convergence and performance of the estimation algorithm.

Since we have some information about the the half-lives of the mRNA and the protein which we use informative independent Gamma priors for both degredation rates v_2 and v_4. All other parameters were chosen to have uniform priors. The performance of the algorithm is very similar to that for the simulated coarse data in that the Markov chains for the main scaled parameters converged well to stationary distributions and, as was the case there, it was not possible to get convergence of the scaling parameters. Figure 6 and Table 2 give estimated posterior distributions and their summary statistics for all other parameters. The estimated posterior densities for the degradation rates only slightly overlap with the prior densities. Both rates are estimated larger than their prior means and similar posteriors are obtained if we use uniform priors. However, the posterior and prior means are still of a similar magnitude and more data would be needed to substantiate the presence of a significant difference between the degradation rates occuring in the two kinds of experiments.

The posterior distribution for the Hill coefficient has a posterior sample median of 8.92 and 95% credibility interval of 5.67 - 15.66. With a probability of 75% or more the value is above 8. Thus these results suggest that the Hill coefficient is quite large and that a substantial amount of real or effective cooperativity is involved in the regulation of the *hes1* gene.

The estimated delay time τ ranges between 22 and 28 minutes with sample mode and median around 25 minutes. This is within a reasonable region of what other authors assumed as delay time (for example, Monk (2003) uses 18 minutes) and is much less than the period of one oscillation.

The Markov chains for the parameters $\tilde{v}_1 = s_M v_1$, $\tilde{k}_1 = s_P k_1$ and $\tilde{v}_3 = (s_P/s_M)v_3$ also have converged well to stationary distributions but since we do not know the scaling factors there are no further interpretations possible about their values other than that which comes from the equality $\tilde{v}_1\tilde{v}_3/\tilde{k}_1 = v_1 v_3/k_1$. The parameter estimates are very similar if we allow for normal measurement error (see Table 2). Plots of the estimated posterior densities can be seen in supplementary Figures 7 - 8.

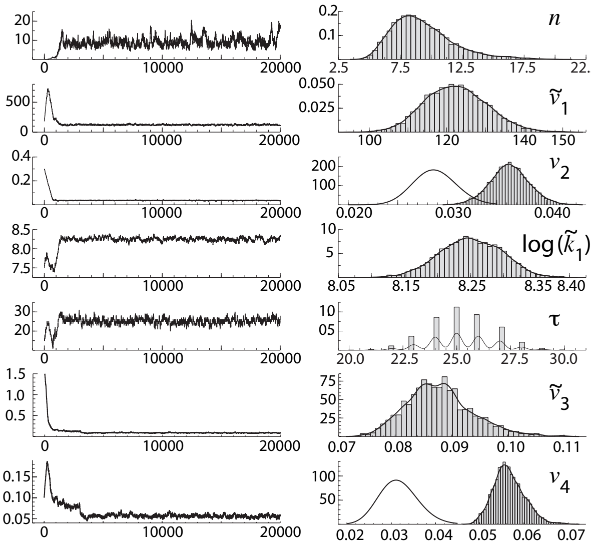

Fig. 6. Chains and posteriors for the Hirata et al. (2002) experimental data. Left: Chains for the various parameters also showing the burn-in. Right: Posterior density estimates for the same parameters (after burn-in). The informative Gamma prior distributions for v_2 and v_4 (solid line) are plotted along with the posterior distributions.

We have, in addition, checked for a wide range of parameter values which have relatively high likelihood that when these are used to simulate data using equations (4) the generated cyclicity is compatible with the observed cyclicity in the experimental data. Indeed, we observe that for these parameters cyclicity is appropriate and robust. Also fitting the distributed delay model to this experimental data gives similar results. These are listed in Table 2 and plots of the posterior distributions are provided in supplementary Figures 5 - 6.

5 DISCUSSION

Our results demonstrate that MCMC methods for SDEs provide practical algorithms for estimating the parameters of simple dynamic regulatory and signalling systems even when the data is as coarse as that considered here. They also indicate the limits of this approach for such data. Although one can only expect to estimate some of the parameters that correspond to the deterministic aspects of the model when the temporal resolution of the data is coarse, with good quality temporally resolved data one can also obtain information about stochastic parameters and population sizes.

The approach presented here has its principles in likelihood based inference, and the estimation process is fully determined by the probabilistic structure of the model. The underlying statistical theory is powerful and well understood and allows us to make statements about the posterior probability of all parameters given the data. This is a major advantage over *ad hoc* methods where the estimation process and model are unrelated. The latter can at best provide point estimates of parameters but lack theoretical justification regarding their properties. Another advantage is that since we are using stochastic models it is possible to use the model in order to probe the nature and magnitude of the stochasticity.

Since almost all biological data that measure the abundance of mRNA and protein are at best only proportional to the true abundance, it is necessary to introduce as unknown parameters the constants of proportionality relating the signal to the abundance for each molecular species. Since the variance of the SDE between times t and $t + \delta t$ is given by the number of events occuring in that time interval, it is proportional to the population size. Therefore, if one can find the scaling parameters, then it is possible to estimate the population sizes of the corresponding molecular species. Another reason that it is necessary to introduce this scaling is the fact that in most cases the data is collected from populations of cells and not from single cells. In this case it is necessary to somehow synchronise the cells so that the average behaviour is similar to the behaviour of individual cells. If this can be done then the scaling parameters correct the model for the fact that it is describing a population of cells. We have shown that these scaling parameters can be estimated if the temporal resolution of the data is sufficiently fine.

The bridging methods we introduced here are a way of reconstructing missing data by probabilistically filling in measurements that were not taken. The major reason for their use here is to be able to perform parameter estimation. More generally it is clear that in a similar spirit MCMC algorithms can be readily constructed to deal with many forms of missing or hidden data. This is likely to be very important for the analysis of regulatory and signalling systems because of the near-impossibility of collecting data on all aspects of the system. Indeed it is likely to be the case that there will only be data on some of the variables that need to be included in a model. We have considered this for some simple models of the *Arabidopsis* circadian clock where only luciferase reporter data are available for some genes and have showed that nevertheless with realistic simulated data it is possible to estimate many scaled parameters effectively (Morton et al. (2005)). It is obvious that latent variables may affect parameter identifiability in the sense that parameters that may be estimated to an acceptable precision for complete data (i.e. when all variables can be measured) may become unidentifiable if one or more variables are not observed. It would be very important to have analytical tools to determine when estimation will be effective with such missing data and when not. The ideas in Rand (2007) and Brown et al. (2004) are a start in this direction.

It seems clear that a very promising avenue for the investigation of regulatory systems is to use these methods with imaging data. Luciferase and gfp reporters combined with fluorescent tagging of protein and mRNA 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 proportional to the abundance of an artificial protein and it is necessary to back-calculate from the time series to the transcription rate. Methods close to those described here can be used for this.

There has been a lot of discussion about whether the oscillations seen in the Hes1 system are due to delay in a network involving a single species or whether they arise from one without delay but involving several species (Hirata et al. 2002; Monk 2003; Jensen et al. 2003; Bernard et al. 2006). In particular, Bernard et al. (2006) recently compared two modelling approaches for the Hes1 system, a simple feedback loop with discrete delay time and a 2-dimensional system including two proteins, Hes1 and Gro/TLE1 incorporating one delay. They conclude in favour of the latter because the additional nonlinear complexity prevents overshooting of the mRNA level and allows for some more realistic fine-tuning of the oscillations. However, we find that when we use the parameter values estimated here there is no problem with such an overshoot and simulations match the data well.

ACKNOWLEDGEMENTS

This research was funded by EPSRC and BBSRC (Interdisciplinary Programme for Cellular Regulation, grant GR/S29256/01) and the EU (BIOSIM Network Contract No. 005137). D. A. Rand is funded by EPSRC Senior Research Fellowship EP/C544587/1. We would like to thank R. Kageyama for providing us with the experimental data and A. Morton for introducing us to computational techniques for bridge building in diffusions.

REFERENCES

M. Barrio, K. Burrage, A. Leier, and T. H. Tian. Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Computational Biology*, 2(9): 1017–1030, 2006. E117.

S. Bernard, B. Čajavec, L. Pujo-Menjouet, M.C. Mackey, and H. Herzel. Modelling transcriptional feedback loops: the role of $Gro/TLE1$ in $Hes1$ oscillations. *Philos Transact Roy. Soc. A Math Phys*, 364(1842):1155–70, 2006.

K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna, and R. A. Cerione. The statistical mechanics of complex signaling networks; nerve growth factor signaling. *Physical Biology*, 1:185–195, 2004.

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.

B. Eraker. MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19:177–191, 2001.

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

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61:781–788, 2005.

A. Golightly and D. J. Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13:838–851, 2006.

Y. Haupt, R. Maya, A. Kazaz, and M. Oren. $Mdm2$ promotes the rapid degradation of $p53$. *Nature*, 387(6630):296–299, 1997.

H. Hirata, S. Yoshiura, T. Ohtsuka, Y. Bessho, T. Harada, K. Yoshikawa, and R. Kageyama. Oscillatory expression of the $bHLH$ factor $Hes1$ regulated by a negative feedback loop. *Science*, 298(5594):840–843, 2002.

A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The $I\kappa B - NF - \kappa B$ signaling module: Temporal control and selective gene activation. *Science*, 298 (5596):1241–1245, 2002.

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. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65:361–393, 1998.

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

M. Koern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews, Genetics*, 6:451–464, 2005.

J. Lewis. Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator. *Current Biology*, 13:1398–1408, 2003.

H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):814–819, 1997.

A. J. Millar, I. A. Carre, C. A. Strayer, N. H. Chua, and S. A. Kay. Circadian clock mutants in Arabidopsis identified by luciferase imaging. *Science*, 267(5201):1161–1163, 1995.

N. A. M. Monk. Oscillatory expression of $Hes1$, $p53$, and $NF - \kappa B$ driven by transcriptional time delays. *Current Biology*, 13(16):1409–1413, 2003.

A. Morton, B. Finkenstädt, E. Heron, and D. A. Rand. Estimation of the arabidopsis circadian clock. *Deliverable D64 of the EU Network of Excellence BioSim: Biosimulation A New Tool in Drug Development*, page 19pp, 2005.

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.

A. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, pages 55–71, 1995.

D. A. Rand. Mapping the global sensitivity of cellular network dynamics. *Submitted*, 2007.

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.