

Inference for a transcriptional stochastic switch model from single cell imaging data

Kirsty Hey^{1*}, Hiroshi Momiji², Karen Featherstone³, Julian Davis³, Mike White⁴,
David Rand², Bärbel Finkenstädt^{1*}

¹ *Department of Statistics, University of Warwick, Coventry, UK*

² *Warwick Systems Biology, University of Warwick, Coventry, UK*

³ *Endocrinology and Diabetes Research Group, University of Manchester, Manchester, UK*

⁴ *Systems Biology Centre, University of Manchester, Manchester, UK*

* K.L.Hey@warwick.ac.uk, B.F.Finkenstadt@warwick.ac.uk

Abstract

Stochastic Reaction Networks (SRNs) can be used to model the temporal behaviour of gene regulation in single cells. In particular, SRNs can capture the features of intrinsic variability arising from the intracellular biochemical processes. However, inference for SRNs is computationally demanding due to the intractability of the transition densities. This paper will show how state space models provide a unifying framework for approximating SRNs with particular attention given to the linear noise approximation (LNA) and an alternative model specific approximation. This methodology has been applied to single cell imaging data measuring expression levels of the human prolactin gene. Transcription is modelled by a random step function relating to bursts in transcriptional activity and we will demonstrate how reversible jump MCMC can be used to infer the switching regimes of this gene within single cells of mammalian tissue. Bayesian Hierarchical; Linear Noise Approximation; Particle Gibbs; Reversible Jump MCMC; Stochastic Reaction Networks; State Space Models; Transcription.

1 Introduction

In individual mammalian cells the expression of some genes, such as prolactin, is highly variable over time, and has been shown to occur in stochastic pulses (Harper et al., 2011). Fluorescent microscopy technology based on the use of reporter genes has proven to be a successful tool for studying the stochastic temporal expression dynamics in individual live cells (Stephens and Allan, 2003). The cell DNA is engineered in such a way that a reporter gene is under the control of the same promoter as the gene of interest so that expression of the reporter can

* To whom correspondence should be addressed.

be used as a proxy for expression of the target gene. Commonly used reporter genes that induce visually identifiable characteristics usually involve fluorescent and luminescent proteins. Examples include the genes for green fluorescent protein (GFP), which cause cells that express it to fluoresce under a laser.

An important statistical problem arising from the use of such reporter constructs is to infer the unobserved transcriptional activity of the reporter which can be assumed to relate to the activity of the gene of interest (Finkenstädt et al., 2008; Harper et al., 2011). At the single cell level, gene expression is fundamentally stochastic (Elowitz et al., 2002; Raj and Van Oudenaarden, 2008) due to both *intrinsic variability* arising from the random microscopic events determining the processes of gene regulation, and *extrinsic variability* between cells caused by both interacting processes and randomness in molecular machinery (Elowitz et al., 2002; Paulsson, 2004). In particular, transcription has been shown to occur in pulsatile bursts for many different genes (Harper et al., 2011; Suter et al., 2011). Here we introduce a general stochastic modelling and inference framework, namely the *stochastic switch model* (SSM), to study gene expression dynamics at single cell level. Our approach is derived on the basis of a stochastic reaction network (SRN) with the assumption that the rate of transcription may change to different levels at unknown times. We estimate the number and temporal location of such switches with a reversible jump Markov chain Monte Carlo (RJMCMC) scheme. As the transcriptional processes cannot be directly observed an essential model ingredient is the measurement equation which links the SSM model to observed imaging time series. The resulting model provides an approach which is both scientifically interpretable and flexible enough to capture a wide range of stochastic dynamics observed in longitudinal single cell imaging data including irregular pulsatile behaviour.

Switch models have previously been considered, for example for inferring transcription factor interactions (Oppen and Sanguinetti, 2010; Sanguinetti et al., 2009) and reconstructing unobserved gene expression dynamics (Finkenstädt et al., 2008; Harper et al., 2011). The general functional form has been a binary model (Harper et al., 2011; Larson et al., 2009; Peccoud and Ycart, 1995; Sanchez et al., 2013; Suter et al., 2011), where transcription can take only two values corresponding to the gene being in an active or inactive state. Although the binary switch model has a simple biological interpretation, restricting transcription to binary states over a prolonged time period may not be flexible enough as other events such as limiting/competing transcription factors may influence gene regulation. Recently, Jenkins et al. (2013) fitted a multiple state switch model to aggregated mRNA populations as observed in microarray analyses and found that the approach is general enough to describe a wide range of observed dynamic patterns in gene expression including oscillatory behaviour with asymmetric cycles of varying amplitude. It is the aim of this study to embed the multiple state switch model within a stochastic reaction network for single cells whilst also introducing a realistic measurement equation with unknown parameters in order to fit the model to experimental single cell imaging time series. Inference is generally challenging due to the intractability of the underlying transition densities of the

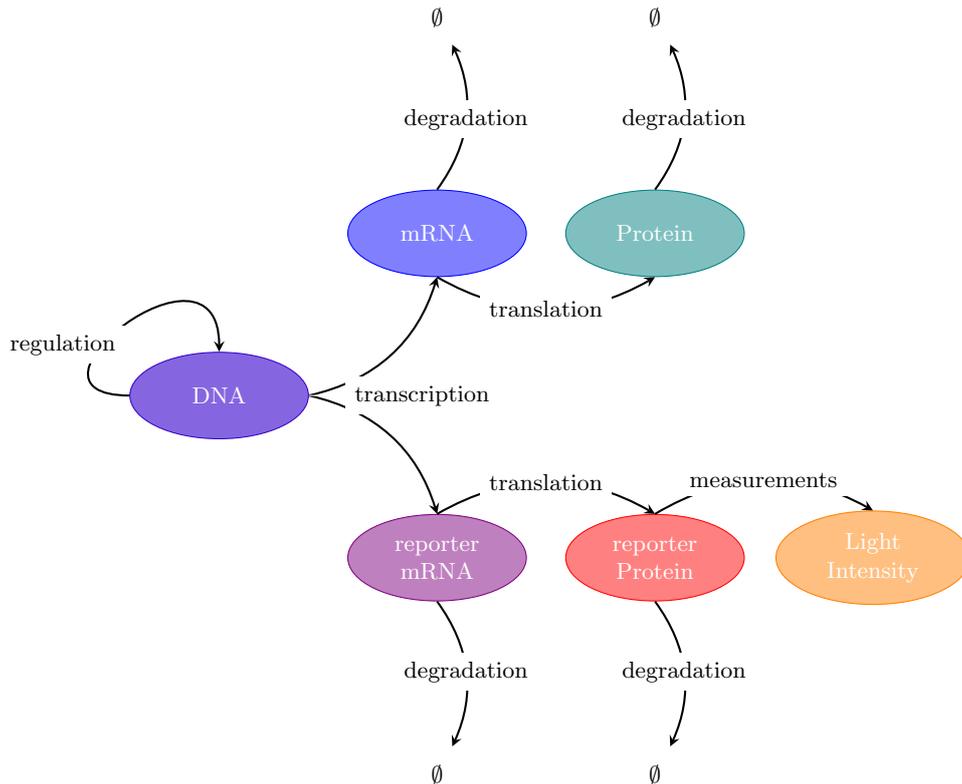


Figure 1: A diagrammatic representation of the transfer of information from DNA to protein through transcription and translation and its relation to the measurement process through a reporter gene construct. Specifically, the cell DNA is engineered such that the reporter is regulated through the same promoter as the gene of interest. Thus transcription of the reporter and native species will be highly correlated. Once transcribed, the mRNA molecules will either degrade or be translated into proteins. Note that there is no longer any coupling between the native and reporter species after transcription and thus the remaining reactions occur with differing rates. The abundance of reporter protein can be measured indirectly through light microscopy techniques.

SRNs. We consider two approximations, namely the linear noise approximation (Komorowski et al., 2009; van Kampen, 1961) and an alternative approximation that is derived specifically for the SSM. We investigate techniques for inference on these approximations and discuss their efficiency and accuracy. We introduce the biological motivation and model in Section 2. A brief overview of SRNs and their associated approximations is given in Section 3 with inferential techniques discussed in Section 4. Section 5 presents results of a simulation study while an application to real data is presented in Section 6. We conclude with a discussion of our findings in Section 7.

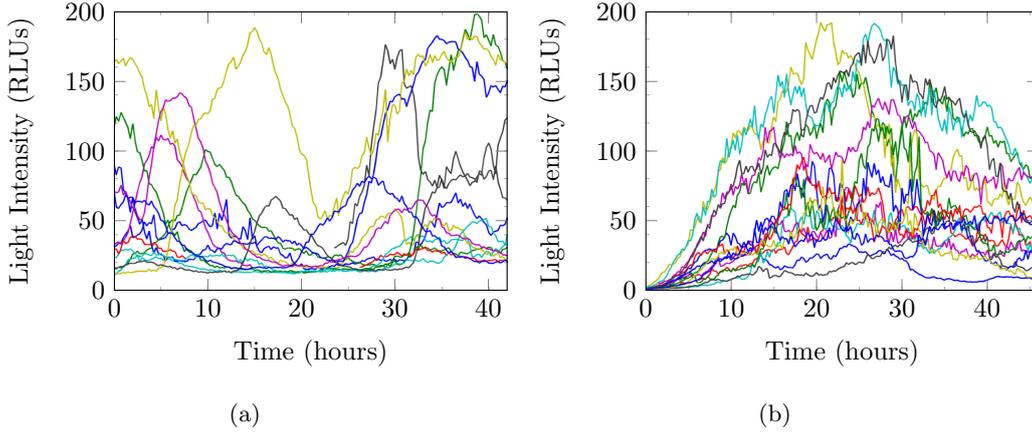


Figure 2: Fluorescent time course data of 15 randomly selected cells from a) an immature rat pituitary tissue slice and b) a mature rat pituitary slice. Measurements were taken every 15 minutes over 42 hours.

2 A stochastic switch model

The basic model of gene expression (Paulsson, 2005) is given by the following four reactions describing the transfer of information encoded within DNA to the creation of protein molecules,



where the superscript for each reaction denotes the corresponding reaction rate. We assume that the transcriptional rate $\beta(t)$ may change over time depending on the activation state of the underlying gene while the rates of translation α and degradation, δ_m and δ_p , are assumed constant in time. Figure 1 gives a diagrammatic representation of how the measurement process, via reporter genes, relates to native gene expression. Consequently, the reporter will be transcribed in parallel to the native gene, creating reporter mRNA which is translated into camera-imageable fluorescent protein probes (see e.g. Elowitz et al. (2002)). The reporter gene also follows reactions (1)-(2) with reporter specific translation and degradation rates. The aim is to estimate the transcriptional dynamics $\beta(t)$ of the reporter, as this will be comparable to the transcriptional dynamics of the target gene. Examples of imaging data are shown in Figure 2. The target gene for these data is the prolactin gene whose regulation is of interest, not only because of the many differing roles it plays within the body, but also its possible link to adenomas of the pituitary (Harper et al., 2010). Figure 2 shows fluorescent time course data for 15 randomly selected cells from tissue samples of an immature as well as a mature rat pituitary slice. The data consist of time discrete observations of the underlying reporter protein levels measured indirectly through the imaging process while both, the reporter protein and mRNA abundances are unobserved.

Following Jenkins et al. (2013), we model the rate of transcription by a piecewise constant function, $\beta(t) = \beta_i$ for $t \in [s_{i-1}, s_i)$, allowing for multiple rates throughout the time course,

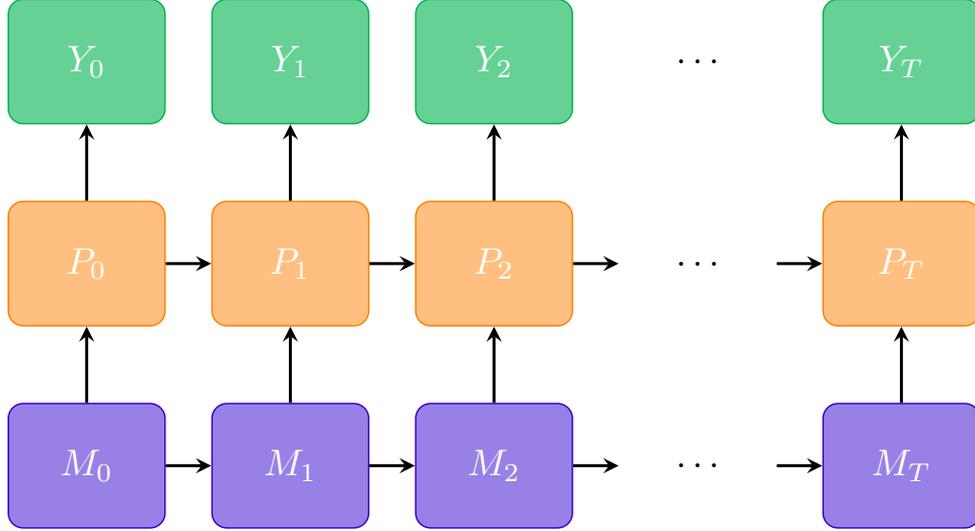


Figure 3: A graphical representation of the hidden Markov model for single cell imaging data. Y_t denotes observed light intensities and are conditionally independent given the latent states M_t and P_t at times t .

where the changes occur at unknown switch times s_1, \dots, s_K . These are likely to be associated with unobserved transcriptional events. We define $s_0 := 0$ and $s_{K+1} = T$, where T is the length of the observed time interval. The observed fluorescent time course together with the latent gene expression dynamics can be modelled by a state space model as depicted in Figure 3 assuming that observations Y_t are conditionally independent given the latent states $\mathbf{X}_t := (M_t, P_t)^T$, consisting of both mRNA (M_t) and protein (P_t), and measurement equation given by,

$$Y_t = \kappa P_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2). \quad (3)$$

The latent states, \mathbf{X}_t , follow a Markov jump process as described in the following section.

3 Stochastic Reaction Networks and Approximations

Stochastic reaction networks can be used to model systems of reactions as in equations (1) and (2) by Markov jump processes (MJPs). Consider a system of ν stochastic reactions involving D molecular species, $X = (X_1, \dots, X_D)^T$ in a well-mixed environment of volume Ω . The stochastic process can be represented by the set of reactions, $\mathcal{P}X \xrightarrow{\mathbf{h}} \mathcal{Q}X$, for matrices \mathcal{P} and \mathcal{Q} . The vector \mathbf{h} , is the vector of hazard functions describing the rate at which each reaction occurs. In general, each hazard function will depend on the state of the system, \mathbf{x} , and the associated kinetic rate of the reaction, denoted by θ . By the law of mass action, the hazard functions are given by,

$$h_j(\mathbf{X}, \theta_j) = \theta_j \prod_{k=1}^D \binom{x_k}{\mathcal{P}_{jk}}, \quad \text{for } j = 1, \dots, \nu,$$

where \mathcal{P}_{jk} is the jk th element of \mathcal{P} and x_k is the k th element of the state vector \mathbf{x} . Given the system is currently in state \mathbf{x} , the probability of reaction j occurring so that the state vector becomes $\mathbf{x} + \mathbf{v}_j$, in the next infinitesimal dt time is given by $h_j(\mathbf{x}, \theta_j)dt$. From this, it is straightforward to derive that the next reaction to occur will be at time $t + \tau$ and of type j with probability,

$$\mathbb{P}(X(t + \tau) = \mathbf{x} + \mathbf{v}_j | \mathbf{X}(t) = \mathbf{x}) = e^{-h_0(\mathbf{x}, \theta)\tau} h_j(\mathbf{x}, \theta_j), \quad (4)$$

where $h_0(\mathbf{x}, \theta) = \sum_{j=1}^{\nu} h_j(\mathbf{x}, \theta_j)$. This identity forms the basis of a stochastic simulation algorithm, (Gillespie, 1977) from which we can generate exact sample paths of a given system. If complete data on all species and reactions were available, inference would be straightforward since an analytical form for the likelihood, using equation (4), is given by

$$f(\mathbf{x}|\theta) = \prod_{i=1}^n h_{j_i}(\mathbf{x}(t_i), \theta_j) \exp\left(-\sum_{i=0}^n h_0(\mathbf{x}(t_i), \theta)[t_{i+1} - t_i]\right), \quad (5)$$

where n is the number of reactions that take place, j_1, \dots, j_n is the sequence of reaction types and t_1, \dots, t_n are the associated timings of each reaction. In molecular biology, complete data paths are not available and, furthermore, commonly only a subset of species are observed, albeit indirectly through a measurement process.

Alternatively, given the above properties, stochastic reaction networks can be formulated as a Markov jump process (MJP) (Stathopoulos and Girolami, 2013), where the D dimensional stochastic process $\mathbf{X}(t) = (X_1(t), \dots, X_D(t))$ satisfies the Markov property,

$$\mathbb{P}(\mathbf{X}(t_i) | \mathbf{X}(t_1), \dots, \mathbf{X}(t_{i-1})) = \mathbb{P}(\mathbf{X}(t_i) | \mathbf{X}(t_{i-1})),$$

for any sequence $t_1 < \dots < t_i$ times. Letting $\mathbb{P}(\mathbf{x}, t) := \mathbb{P}(\mathbf{X}(t) = \mathbf{x} | \mathbf{X}(0) = \mathbf{x}_0)$ denote the transition density from time 0 to time t , it can be shown to satisfy Kolmogorov's forward equations from which the following, often intractable, ordinary differential equation (ODE), termed *chemical master equation*, can be obtained,

$$\frac{d}{dt} \mathbb{P}(\mathbf{x}, t) = \sum_{j=1}^J h_j(\mathbf{x} - \mathbf{v}_j, \theta) \mathbb{P}(\mathbf{x} - \mathbf{v}_j, t) - h_j(\mathbf{x}, \theta_j) \mathbb{P}(\mathbf{x}, t), \quad (6)$$

$$\mathbb{P}(\mathbf{x}, 0) = \mathbb{I}[\mathbf{x} = \mathbf{x}_0]. \quad (7)$$

One approach for performing inference on SRNs is to work with the complete system (5) achieved through integrating over the latent reaction paths and attention has been focused on performing these high-dimensional integrations in a computationally efficient way. For example, Andrieu et al. (2009) show how an application of particle MCMC methods can be used to perform inference on MJPs, in particular the stochastic kinetic Lotka-Volterra model. This was found to be computationally burdensome (Golightly and Wilkinson, 2011) and the method was shown to perform poorly in low measurement error scenarios. Other approaches to performing inference on the exact system include a reversible jump MCMC method (Boys et al., 2008) and

the MCEM² (Monte Carlo Expectation Maximisation with modified cross-entropy method) of Daigle et al. (2012) which makes use of rare simulation techniques. Amrein and Künsch (2012) develop a technique to sample hidden trajectories conditional on the observed endpoints when species are observed with unknown measurement error but known scaling. All the methods discussed here have so far only been applied to synthetic data. One exception is the delayed acceptance MCMC method of Golightly et al. (2014) who apply their methodology to an epidemic dataset. In addition, the majority of these papers assume that one can directly observe state variables. This is unrealistic within a molecular biology framework as experimental methods will invariably result in a measurement process with error and scaling as the direct number of molecular populations are not observable. The incorporation of both unknown measurement error and scaling is non-trivial in this framework and we consider it in detail in the following. In contrast to the above, we consider the feasibility of approximating the underlying MJP. We first discuss approximations common in the literature and then derive a further, model specific, approximation. The reader is referred to Appendix A for details of each derivation.

In the macroscopic limit, a deterministic approximation, is derived by taking the conditional expectation of the stochastic MJP given its history (Chesson, 1978). The resulting approximation, \mathbf{X}^D , satisfies the following ODE,

$$\frac{d\mathbf{X}^D}{dt} = A(\mathbf{X}^D) := \sum_{j=1}^{\nu} \mathbf{v}_j h_j(\mathbf{X}^D, \theta) = S\mathbf{h}(\mathbf{X}^D, \theta), \quad \mathbf{X}^D(0) = \mathbf{x}_0. \quad (8)$$

Clearly, all stochasticity is removed by this approximation and is therefore appropriate for modelling aggregate data with no intrinsic variability as in Jenkins et al. (2013).

There are two mesoscopic approximations in the literature that have been used to model SRNs which give rise to systems of stochastic differential equations. These are the chemical Langevin equation (CLE), and the linear noise approximation (LNA). The CLE was first derived heuristically in Gillespie (2000), with more rigorous treatment given in Anderson and Kurtz (2011). The derivation relies on two so-called *leap conditions*, the first that the time between observations is small enough that the hazard rates can be considered constant and the second, that the rescaled concentration hazard rates are large. The chemical Langevin approximation has been used extensively for inference within SRNs, (Golightly and Wilkinson, 2005, 2011; Heron et al., 2007). Despite this there are several drawbacks to using this approximation, not least of all, that the transition density often remains intractable. Moreover, in practice, data are measured at discrete time intervals that cannot be assumed to satisfy the first leap condition.

The linear noise approximation (LNA), on the other hand, is a linearisation of the master equation and always results in analytical transition densities. Derivations of varying degrees of rigour can be found, for example see Kurtz (1971); van Kampen (1961) and Wallace et al. (2012). In particular, Kurtz (1971) show how it can be viewed as a central limit theorem for

the underlying MJP, as the system size $\Omega \rightarrow \infty$. The full specification of the LNA is given by,

$$\mathbf{X}^L(t) = \phi(t) + \Omega^{-1/2}\xi(t), \quad (9)$$

$$\frac{d\phi}{dt} = A(\phi) \quad (10)$$

$$d\xi = J(\phi(t))\xi dt + \sqrt{B(\phi)}dW_t, \quad (11)$$

where ϕ is the macroscopic ODE solution, dW_t are independent Wiener processes, A is defined as in (8), $B := S \text{diag}(\mathbf{h}(\phi, \theta))S^T$ and $J = (J_{ij}) = \left(\frac{\partial A_j}{\partial \phi_j}\right)$ is the Jacobian. Equation (11) is linear with Itô representation and thus the transition $\mathbb{P}(\xi(t + \tau)|\xi(t))$ is Gaussian with mean and variance defined by (Komorowski et al., 2009),

$$\frac{d\mu}{dt} = J(t)\mu(t) \quad (12)$$

$$\frac{d\Sigma}{dt} = \Sigma(t)J(t)^T + J(t)\Sigma(t)^T + B(t)B(t)^T. \quad (13)$$

Thus, the transition probabilities of the state vector are given by, $\mathbb{P}(\mathbf{X}^L(t+\tau)|\mathbf{X}^L(t)) = N(\phi(t) + \Omega^{-1/2}\mu(t + \tau), \Omega^{-1}\Sigma(t + \tau))$, (see Finkenstädt et al. (2013)). In the case of a linear system, as in the gene transcription model, (12) and (13) can be simplified to give, $\mu(t + \tau) = e^{J\tau}\xi(t)$ and $\Sigma(t + \tau) = \int_t^{t+\tau} [e^{J(t+\tau-s)}B(s)][e^{J(t+\tau-s)}B(s)]^T ds$.

Both the CLE and the LNA are derived in the limit as the system size $\Omega \rightarrow \infty$. For precise statements of the limits used in this derivation see Kurtz (1971, 1978). It is common in the literature for the LNA to be derived as an approximation to the CLE which can lead to the erroneous assumption that the LNA also only holds under the two leap conditions. Anderson and Kurtz (2011) show that this is not the case and less stringent assumptions are required (Appendix A). Inference on transcription networks based on the LNA can be found in Finkenstädt et al. (2013) for a constant transcription model, Komorowski et al. (2009) for a nonlinear autoregulatory system and Fearnhead et al. (2014) for a dimerisation system. Although the LNA is derived in the limit as the system size gets large, these studies have found that reasonable estimation can be performed even when the system size is of a mesoscopic size. In addition, Fearnhead et al. (2014) introduce a variant, the restarting LNA, where the initial condition of the ODE for ϕ , is updated depending on the previous state. This becomes important to limit the effect of the initial value for the LNA system.

In addition, the gene transcription model (1)-(2) gives rise to a further model-specific approximation. This is derived from the factorisation,

$$\begin{aligned} \mathbb{P}(M_t, P_t|M_{t-1}, P_{t-1}) &= \mathbb{P}(M_t|M_{t-1}, P_{t-1}) \mathbb{P}(P_t|M_{t-1}, M_t, P_{t-1}) \\ &\approx \mathbb{P}(M_t|M_{t-1}) \mathbb{P}(P_t|M_t, P_{t-1}), \end{aligned} \quad (14)$$

from which one can construct two separable SRNs,



where $M^*(t) := m_t$, the current mRNA population. Under this birth-death decomposition, one can obtain the exact transition densities for the two separable birth-death subsystems in (15)-(16) (Appendix A.4). The resulting transition densities are a convolution of Poisson and binomial random variables which may be approximated, for example, by,

$$\begin{aligned} M_t | M_{t-1} = m_{t-1} &\sim N_T(\lambda_t^m + m_{t-1}\pi_t^m, \lambda_t^m + m_{t-1}\pi_t^m(1 - \pi_t^m)), \\ P_t | M_t = m_{t-1}, P_{t-1} = p_{t-1} &\sim N_T(\lambda_t^p + p_{t-1}\pi_t^p, \lambda_t^p + p_{t-1}\pi_t^p(1 - \pi_t^p)), \end{aligned}$$

where N_T denotes the normal density truncated to the positive real line and $\lambda^m, \pi^m, \lambda^p, \pi^p$ satisfy the following system of ODEs,

$$\begin{aligned} \frac{d\lambda^m}{dt} &= \beta(t) - \delta_m \lambda^m(t), & \frac{d\pi^m}{dt} &= -\delta_m \pi^m(t), & \lambda^m(0) &= 0, & \pi^m(0) &= 1. \\ \frac{d\lambda^p}{dt} &= \alpha m_t - \delta_p \lambda^p(t), & \frac{d\pi^p}{dt} &= -\delta_p \pi^p(t), & \lambda^p(0) &= 0, & \pi^p(0) &= 1. \end{aligned}$$

We term this approach the birth-death approximation (BDA) and to justify its use, we present the following validation study. First, note that the approximation only affects inference regarding the protein process since marginal inference for the mRNA process will be exact. Figure 4 shows how the transition density behaves under the true process, the LNA and the BDA for 3 different scenarios. It can be seen that the discrete BDA matches very closely to the true density in all scenarios and moreover the truncated normal approximation models the skewness at low molecular numbers better than the symmetric LNA. The improved precision of the BDA becomes apparent in Figure 5, which shows the 95% pointwise confidence envelopes for processes simulated from the various approximations. In all scenarios, the BDA envelopes for both mRNA and protein are closer to the true envelopes than either the LNA or the restarting LNA. The restarting LNA improves as molecular numbers increase although consistently overestimates the variance for low numbers and consequently will be likely to miss switch points in the transcriptional profiles. This empirical validation supports the intuition that the BDA may be a preferable approximation to the more standard LNA especially at low molecular levels.

4 Inference

State space models provide a unifying framework for modelling SRNs through any of the approximations described above, since in the presence of a measurement process, we have that,

$$X_{t+1} \sim h(x_{t+1}|x_t, \theta) \tag{17}$$

$$Y_t \sim g(y_t|x_t, \theta), \tag{18}$$

where h is the transition density of the approximating SRN (Appendix B) and g is the density of the measurement process (e.g. equation (3)). For ease of notation, we have dropped any explicit dependence on time, i.e. the sequence of observations (Y_0, \dots, Y_T) are assumed to occur at arbitrary times, $(0, t_1, \dots, T)$ and are equivalent to $(Y(0), Y(t_1), \dots, Y(T))$ in the notation of

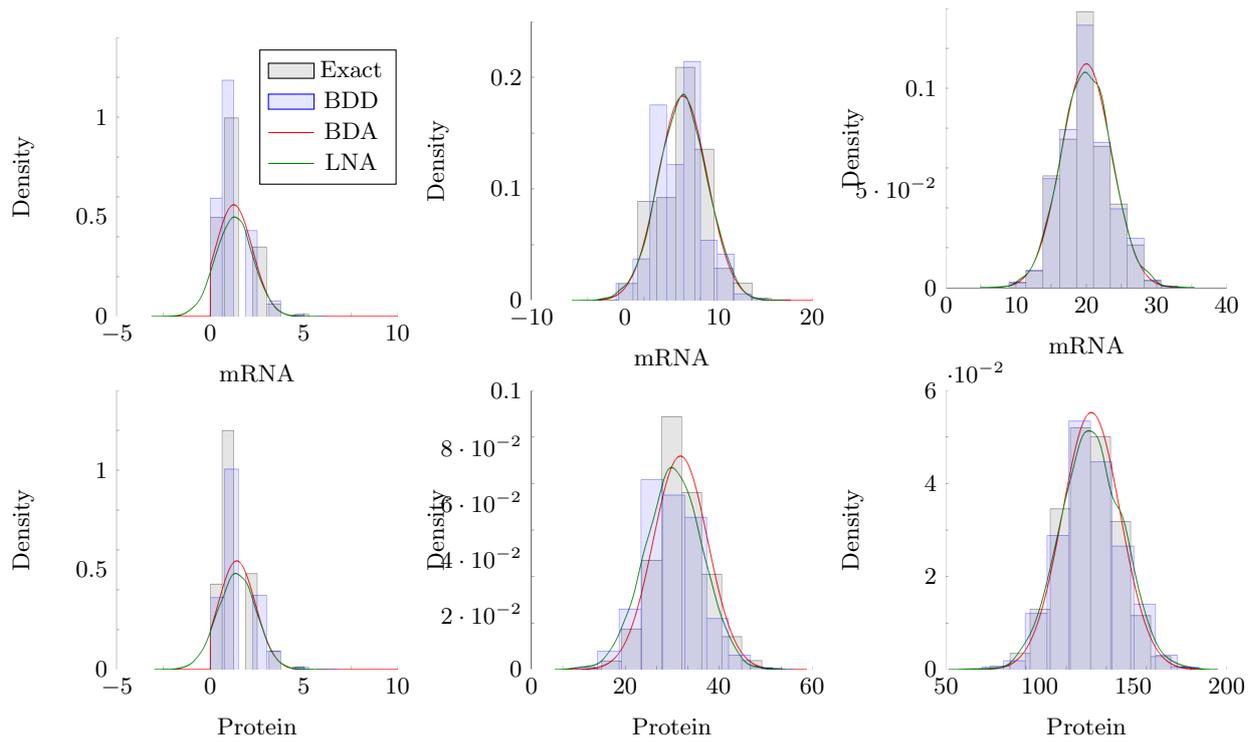


Figure 4: Transition densities calculated for the exact process (obtain empirically and shown in grey), the BDA (blue), the normal approximation to the BDA (red) and the LNA (green). These have been calculated for 3 different scenarios for both mRNA (top) and for Protein (bottom). Scenario 1 (left) is calculated with parameters $\delta_m = 1, \delta_p = 0.7, \alpha = 1, \beta = 1$ over a time interval of 0.5 (hours). Scenario 2 (centre) is calculated with parameters $\delta_m = 1, \delta_p = 0.7, \alpha = 3, \beta = 6$ over a time interval of 0.5 (hours). Scenario 3 (right) is simulated with parameters $\delta_m = 1, \delta_p = 0.7, \alpha = 3, \beta = 20$ over a time interval of 0.5 (hours).

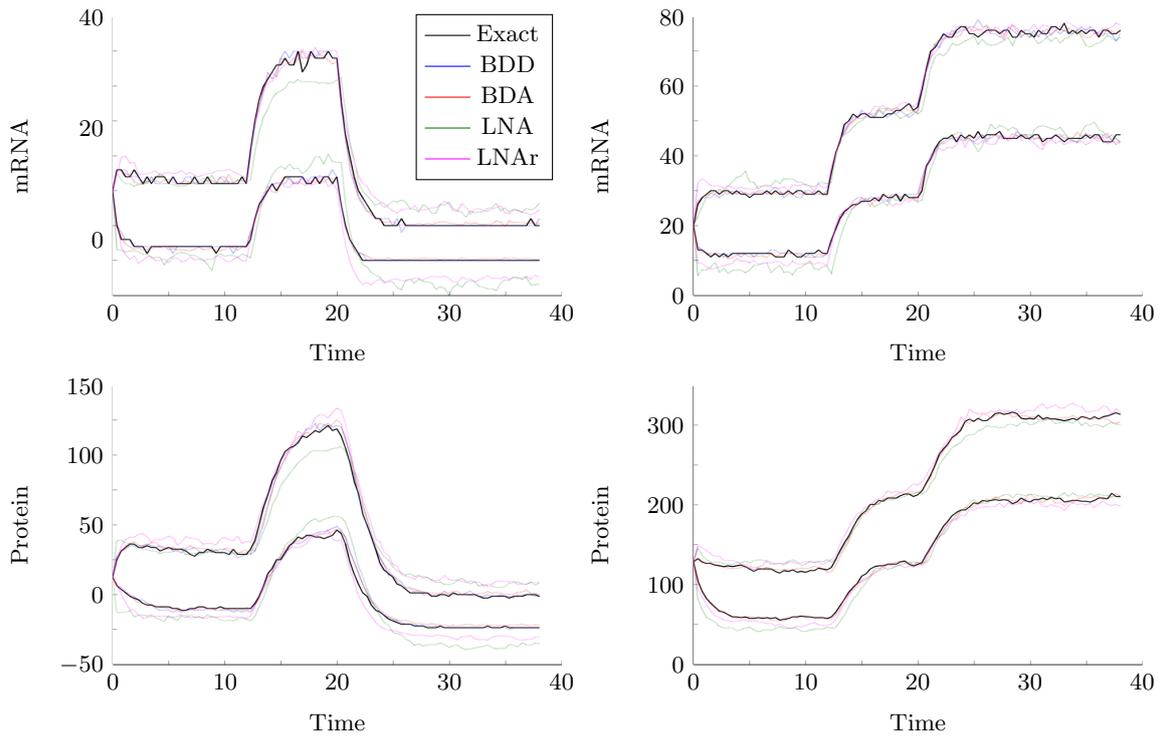


Figure 5: 95% pointwise confidence envelopes for simulated mRNA (top) and protein (bottom) processes under the true process (black), the BDA (blue), the normal approximation to the BDA (red), the LNA (green) and the restarting LNA (pink) for two different scenarios. Scenario 1 (left) is simulated from the parameters $\delta_m = 1, \delta_p = 0.7, \alpha = 3, \beta_0 = 6, \beta_1 = 20, \beta_2 = 2$ with switches occurring at $t = 12, 20$. Scenario 2 (right) is simulated with parameters $\delta_m = 1, \delta_p = 0.7, \alpha = 3, \beta_0 = 20, \beta_1 = 40, \beta_2 = 60$ with switches occurring at $t = 12, 20$

Section 3. We now investigate the performance of two approximations to the underlying SRN, namely the LNA and the BDA neglecting the Langevin approximation due to the theoretical arguments given in Anderson and Kurtz (2011). For the remainder of this paper, f will be used to denote a density whose meaning will be clear from the arguments and θ will denote the vector of unknown parameters. Letting t range from $0, \dots, T$ we will use the notation $y_{0:t}$ to define the sequence of observations y_0, \dots, y_t . In addition, we will use the convention $\mathbf{y} := y_{0:T}$ to denote all the observed data points y_0, \dots, y_T . The data likelihood is given by the marginal density,

$$f(\mathbf{y}|\theta) = \int_{\mathbf{x}} f(\mathbf{y}, \mathbf{x}|\theta) d\mathbf{x}, \quad (19)$$

$$= \int_{\mathbf{x}} h(x_0|\theta)g(y_0|x_0, \theta) \prod_{t=1}^T h(x_t|x_{t-1}, \theta)g(y_t|x_t, \theta) d\mathbf{x}. \quad (20)$$

If the integral in equation (19) cannot be computed explicitly, for example under the BDA, one can use a two step Gibbs sampler (GS) to perform inference on the parameter vector θ ,

1. Sample the parameter vector θ from $f(\theta|\mathbf{y}, \mathbf{x})$.
2. Sample the latent states, \mathbf{x} , from the filtering density, $f(\mathbf{x}|\mathbf{y}, \theta)$.

On the other hand, when (19) can be evaluated explicitly, inference about θ can be achieved by sampling θ from the posterior $f(\theta|\mathbf{y})$. This is the case under the LNA with Gaussian measurement error, where one can use the Kalman methodology to evaluate (19), details of which can be found in Appendix C.

4.1 Parameter Inference

In order to sample θ from either $f(\theta|\mathbf{y}, \mathbf{x})$ or from the full posterior $f(\theta|\mathbf{y})$, depending upon which approximation is used, we construct an appropriate MCMC sampler targeting the correct posterior density. In particular, inference about θ includes inference on the number, k , and position, s_1, \dots, s_k of switches as well as the associated kinetic parameters $(\beta_0, \dots, \beta_k, \alpha, \delta_m, \delta_p)$, the measurement parameters, $(\kappa, \sigma_\epsilon^2)$ and the initial state of the latent molecular processes, (M_0, P_0) . Note that the dimension of θ varies with the number of switches k and thus the MCMC sampler is required to sample across these dimensions. In order to achieve this, we employ a reversible jump scheme (Green, 1995). This is implemented according to a similar specification as in Jenkins et al. (2013). Specifically, we define a Poisson distribution over the prior number of switch points within the data conditional on the number of switches not exceeding some k_{\max} . It is assumed that the prior position of switch points is uniform on the entire observation window, $[0, T)$. In order to sample across the differing dimensions at each iteration of the MCMC, we propose one of three possible moves, 1) add a switch, 2) delete a switch and 3) move a switch. Further details of these proposals are given in Appendix E. For the kinetic parameters, these are updated via a random walk Metropolis-Hastings (MH) step, where we impose log-normal priors to account for the non-negativity and possible skewness of the parameters.

Due to the high dimensionality of the integral in (19), there is a strong correlation between different model parameters and thus it is desirable to sample parameters efficiently. It was found that targeting the posterior of the log-parameters ($\log \theta$) resulted in a more efficient MCMC scheme as both the proposals and priors are symmetric in log-space. Efficiency was further increased by implementing the adaptive MCMC scheme of Haario et al. (2001). The log-parameters are sampled in two blocks where proposals are drawn from a multivariate normal distribution centred at the previous value and with covariance matrix proportional to the covariance of the Markov chains up until that iteration. This adaptation results in an ergodic Markov chain provided the target density is bounded from above and has a bounded support.

4.2 Inferring the latent states

There are many ways one can perform the filtering procedure in step 2 of the Gibbs sampler each of which are subject to pros and cons, in particular the computational burden (see, Fearnhead (2011) for a review). We have found that a conditional sequential Monte Carlo particle filter (Andrieu et al., 2009, 2010) performs well under the BDA. The sequential Monte Carlo approach is based on forward simulations to sequentially approximate the filtering density, $f(x_t|y_{0:t}, \theta)$ and can be applied to very general state space models that are not necessarily linear or Gaussian. More specifically, the filtering density is approximated by, $f^N(x_t|y_{0:t}, \theta) = \sum_{i=1}^N w_t^{(i)} \delta_{x_t^{(i)}}$, where δ_x is a delta function centred at x , and $w_t^{(i)}$ are importance weights. Two steps are needed to obtain a sample $\{x_t^{(i)}, w_t^{(i)}\}$:

1. Sample $x_t^{(i)} \sim q(\cdot|x_{t-1}^{(i)}, y_{0:t})$, where q is the importance density.
2. Compute the importance weights.

Given the approximate filtering density $f^N(\mathbf{x}|\mathbf{y}, \theta) := f^N(x_{0:T}|y_{0:T}, \theta)$, one can obtain a sample of the latent states \mathbf{x} as required in the two step Gibbs procedure (GS) and the resulting algorithm is termed *Particle Gibbs* (Andrieu et al., 2009, 2010). Further details of the algorithm and proposal densities used for the BDA can be found in Appendix D.

4.3 Hierarchical Modelling

In order to incorporate as much information as possible into the algorithm, informative prior distributions are desirable. In the example of single cell imaging data, additional experiments can be performed to obtain estimates of the two degradation parameters, δ_m and δ_p , however, little is known about the remainder of the parameters. Typically, a dataset will consist of multiple time series from the same biological conditions. As in Finkenstädt et al. (2013) one can assume that parameters will be similar between cells and come from some overall population which will be incorporated through a hierarchical structure. This aids in parameter identifiability as strength is borrowed across the multiple time courses (Gelman et al., 2013).

Let $\mathbf{y}^{(i)}$ denote the observed time series for cell i , and $\theta^{(i)} := (\beta^{(i)}(t), \alpha^{(i)}, \delta_m^{(i)}, \delta_p^{(i)}, \kappa^{(i)}, \sigma_\epsilon^{(i)}, M_0^{(i)}, P_0^{(i)})$, the vector of parameters, for $i = 1, \dots, N$. We assume a hierarchical structure for the translation rates, $\log \alpha^{(i)} \sim N(\mu_\alpha, \sigma_\alpha^2)$, and the measurement parameters, $\log \kappa^{(i)} \sim N(\mu_\kappa, \sigma_\kappa^2)$, $\log \sigma_\epsilon^{(i)} \sim N(\mu_\sigma, \sigma_\sigma^2)$. In contrast to the standard gamma priors, the log-normal specification allows for a conjugate update of the hyper-parameters, which is a distinct practical advantage in terms of computational efficiency.

Specifying a hierarchical model for the transcription rates $\beta := (\beta^{(1)}, \dots, \beta^{(N)})$, where $\beta^{(i)} := (\beta_0^{(i)}, \dots, \beta_K^{(i)})$ is the vector of rates for each cell $i = 1, \dots, N$, is less straightforward. To use the same specification as above would dilute the effect of switching events since all rates would be shrunk to a single tighter distribution. On the other hand, vague proper priors are not a feasible option since it gives too much prior probability to the zero switch model (Berger, 2000; Green, 1995). As an alternative, we specify a hierarchical mixture model where, $\log \beta^{(i)} \sim \sum_{m=1}^M w_{\beta_m} N(\mu_{\beta_m}, \sigma_{\beta_m}^2)$. Allowing a mixture model reduces the shrinkage due to the hierarchical specification. However, without resorting to a second reversible jump scheme, it is necessary to specify the number of components in advance. One could chose several candidates and perform model selection *a posteriori*, although we have found that the use of two components is sufficient to capture the variability in the data and is supported by the biological hypothesis that transcription will typically occur at either a high or low rate. Results from simulations show that if the rates truly come from a single component mixture model then this is elicited from a two component specification with one very low weight estimated.

The hyper-parameters $\vartheta := (\mu_\alpha, \sigma_\alpha^2, \mu_\kappa, \sigma_\kappa^2, \mu_\sigma, \sigma_\sigma^2, \mu_\beta, \sigma_\beta^2, \mathbf{w}_\beta)$ are estimated in addition to each $\theta^{(i)}$, $i = 1, \dots, N$. We specified uninformative priors for each of the hyper-parameters where the mean was given a $N(0, 100^2)$ prior, and the precision was given a $\text{Gamma}(1, 0.001)$ prior while the weights of the hierarchical mixture model have a $\text{Dirichlet}(2, 2)$ prior.

In order to draw samples from the full posterior $f(\theta^{(1)}, \dots, \theta^{(N)}, \vartheta | y^{(1)}, \dots, y^{(N)})$, we iterate sampling the hyper-parameters from the full conditional,

$$f(\vartheta | \theta^{(1)}, \dots, \theta^{(N)}, y^{(1)}, \dots, y^{(N)}) = f(\vartheta | \theta^{(1)}, \dots, \theta^{(N)})$$

(details of the conjugate specification are given in Appendix F) with sampling the parameter estimates given the hyper-parameters and observed data,

$$f(\theta^{(1)}, \dots, \theta^{(N)} | y^{(1)}, \dots, y^{(N)}, \vartheta) = \prod_{i=1}^N f(\theta^{(i)} | y^{(i)}, \vartheta)$$

. Since all $\theta^{(i)}$ are independent given the hyper-parameters and data, we simplify step this second step to sample from each $f(\theta^{(i)} | y^{(i)}, \vartheta)$ in turn by the MCMC sampler described in Section 4.1. There is clearly a trade-off between the number of cells to update hierarchically and the computational cost involved, which will be investigated in the following.

5 Simulation Study

In order to investigate the performance of the LNA and the BDA we perform a comprehensive simulation study where synthetic data were generated from the exact Markov jump process via a stochastic simulation algorithm (Gillespie, 1977) based on (4). In particular, data were constructed to replicate the main features of the observed data in Figure 2. We consider 4 scenarios of different parameter choices relating to 4 different underlying population levels where each dataset contains 15 time series each measured over 30 hours with 100 discrete measurements. Within each scenario, time series are simulated with a variety of switching regimes and for each scenario we performed 10 simulations and applied both the BDA and LNA models. In both cases it was found that using informative priors for the degradation parameters was essential in order to identify both the transcriptional profile, $\beta(t)$, and the translation rate, α . We therefore imposed informative prior distributions, $\log \delta_m \sim N(\mu_{\delta_m}, \sigma_{\delta_m}^2)$ and $\log \delta_p \sim N(\mu_{\delta_p}, \sigma_{\delta_p}^2)$ where μ_{δ_m} , μ_{δ_p} , $\sigma_{\delta_m}^2$ and $\sigma_{\delta_p}^2$ were all fixed at the true values.

Preliminary analyses showed that under the BDA, the scaling parameter, κ , remained unidentifiable in the majority of simulations. We hypothesise that this is due to the fact that under the BDA, we are targeting an extended space by explicitly sampling the latent states. To our knowledge, there has been no application within this extended framework that has been able to incorporate a scaling parameter in the measurement equation. We hence consider two scenarios, (1) κ is fixed at the true value and (2) κ is fixed at the posterior median obtained from the LNA. The motivation for this becomes clear in the following.

As stated above, our sampler targets the log-posterior of the rate parameters which greatly increases the efficiency and also allows the use of a conjugate update of the hyper parameters. Despite the hierarchical setup and the informative priors, the sampler still suffered from poor mixing and it was found, as in Finkenstädt et al. (2013), that re-parameterising the model in terms of $\tilde{\alpha} := \kappa\alpha$ and $\tilde{P}_t := \kappa P_t$ greatly improved the mixing properties of the Markov chains. The algorithm specification for the LNA and BDA models are as follows:

Case LNA: Under the LNA, inference is performed through the following MCMC sampler,

- (1) Initialise parameters.
- (2) Update model parameters.
 - a) Sample hyper parameters from the full conditional distribution.
 - b) Update the log transcriptional step function by RJ step.
 - c) Sample $\log(\beta_0, \dots, \beta_k, \delta_m, \delta_p, M_0)$ parameters by a random walk MH step.
 - d) Sample $\log(\tilde{\alpha}, \kappa, \sigma_c^2, \tilde{P}_0)$ parameters by a random walk MH step.
- (3) Repeat step 2 until convergence.

Case BDA: Under the BDA, inference is performed in the following sampler,

- (1) Initialise model parameters, θ , and latent states $(M_1, \dots, M_T, \tilde{P}_1, \dots, \tilde{P}_T)$
- (2) Update model parameters.

- a) Sample hyper parameters from the full conditional distribution.
 - b) Update transcriptional step function by RJ step.
 - c) Sample $\log(\beta_0, \dots, \beta_k, \delta_m, \delta_p, M_0)$ parameters by a random walk MH step.
 - d) Sample $\log(\tilde{\alpha}, \sigma_\epsilon^2, \tilde{P}_0)$ parameters by a random walk MH step.
- (3) Update the latent states, $M_1, \dots, M_T, \tilde{P}_1, \dots, \tilde{P}_T$, by particle Gibbs.
- (4) Repeat steps 2 and 3 until convergence.

5.1 Results

The simulation study was coded in MATLAB[®] (The MathWorks Inc., 2012) and typically under the LNA, we found 300K iterations to be sufficient to achieve convergence with a burn in period of 75K. Despite the fact that the BDA methodology is computationally faster to run per MCMC iteration, due to the higher autocorrelation in the chains and poorer mixing properties, we found it that it could take anywhere between 300K and 1million iterations to sufficiently explore the posterior. This is unsurprising since, the BDA methodology requires the sampling of all the latent states in addition to the parameter vector. For all scenarios under the BDA, 100 particles were used to give a sufficient number of independent samples in the particle filter.

The key parameter of interest will typically be the estimated transcriptional profiles for $\beta(t)$. To obtain the posterior profiles, we extract the marginal distribution of the number and positioning of switch times. Conditional on these times, the posterior rates are then obtained from the MCMC output. Figure 6 shows the mean square errors (MSE) of each parameter calculated at the posterior median values for each of the 4 scenarios. As can be expected the BDA (red diamonds) has lower MSE for the parameters $\beta(t)$ and α than the LNA (blue diamonds). This may be because the BDA is estimating one fewer parameters with κ fixed at the true value. However, the increased accuracy remains also when κ is fixed at the truth under the LNA (blue squares). Since in general one may not know the value of κ *a priori*, we have also run the BDA model with κ fixed at the posterior median calculated under the LNA. The MSE for these runs (red squares) shows that the estimated transcriptional profile in particular remains more accurate than when running the LNA alone. Having said this, Figure 7 shows the corresponding widths of the 50% credible intervals for each parameter which are generally comparable between the LNA and BDA with the exception of scenario 2, where the precision of the BDA estimate is less than that of the LNA. A possible alternative would be to run the BDA over a grid of “reasonable estimates” for κ and perform model selection.

6 Application to Data

We apply both the LNA and BDA methodologies for calculating the underlying transcriptional profiles for the experimental data in Figure 2. We first apply the LNA and then apply the BDA with κ fixed at the posterior median obtained from the LNA. When applying to real

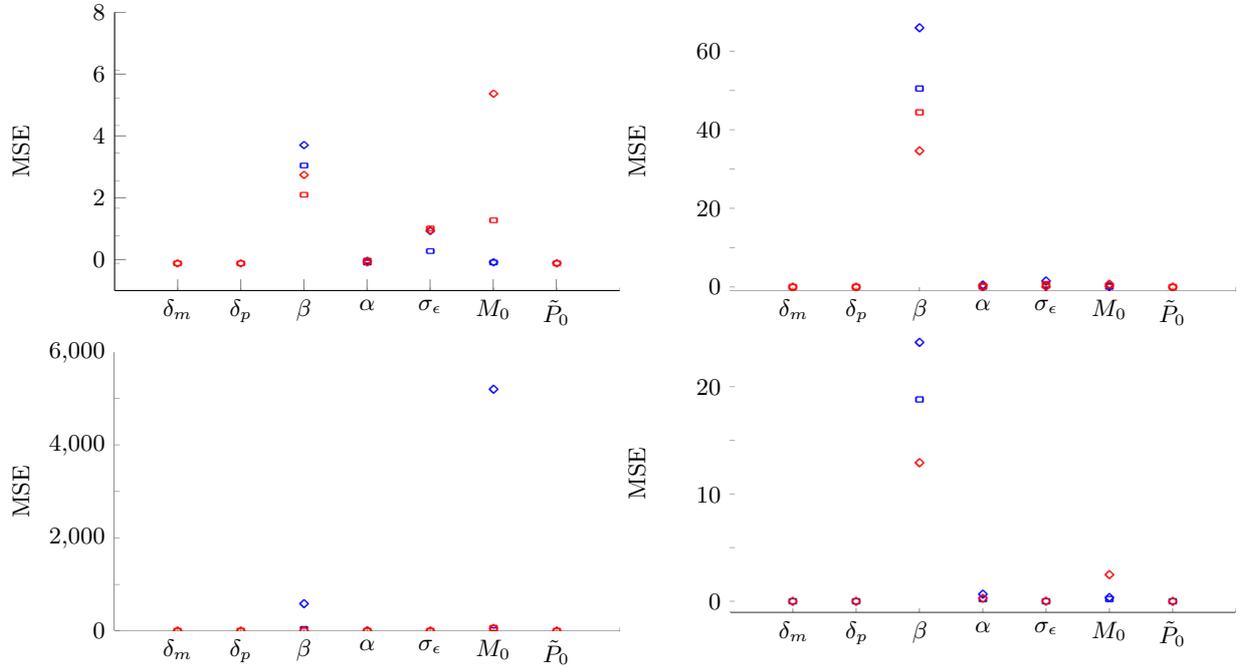


Figure 6: The mean square error for each estimated parameter calculated under the LNA (blue diamond), the LNA with κ fixed at the truth (blue square), the BDA with κ fixed at the truth (red diamond) and the BDA with κ fixed at the posterior median of the LNA. Each panel corresponds to a different set of parameters for systems of different molecular numbers. For each scenario, we ran 10 different simulations each of which contained 15 individual time series. The MSE is therefore calculated from 150 different parameter estimates. Specifically, we drew the parameter sets from the hyper-distributions. Scenario 1 (top left) is simulated from the set of hyper-distributions: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim N(\log(8), 0.2)$, $\log \alpha \sim N(\log(4), 0.05)$, $\log \kappa \sim N(\log(2), 0.05)$, $\log \sigma_\epsilon \sim N(\log(4), 0.2)$. Scenario 2 (top right) is simulated from the set of hyper-distributions: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim 0.5 * N(\log(2), 0.2) + 0.5 * N(\log(10), 0.1)$, $\log \alpha \sim N(\log(4), 0.05)$, $\log \kappa \sim N(\log(2), 0.05)$, $\log \sigma_\epsilon \sim N(\log(4), 0.2)$. Scenario 3 (bottom left) is simulated from the set of hyper-distributions: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim 0.5 * N(\log(2), 0.2) + 0.5 * N(\log(4), 0.1)$, $\log \alpha \sim N(\log(1), 0.05)$, $\log \kappa \sim N(\log(4), 0.05)$, $\log \sigma_\epsilon \sim N(\log(5), 0.2)$. Scenario 4 (bottom right) is simulated from the set of hyper-distributions: $\log \delta_m \sim N(\log(0.4), 0.02)$, $\log \delta_p \sim N(\log(0.7), 0.02)$, $\log \beta \sim 0.5 * N(\log(20), 0.2) + 0.5 * N(\log(40), 0.1)$, $\log \alpha \sim N(\log(1), 0.05)$, $\log \kappa \sim N(\log(2), 0.05)$, $\log \sigma_\epsilon \sim N(\log(15), 0.2)$. In all cases we allowed up to 3 switches to occur in transcription. The MSE is calculated at the posterior median of each parameter.

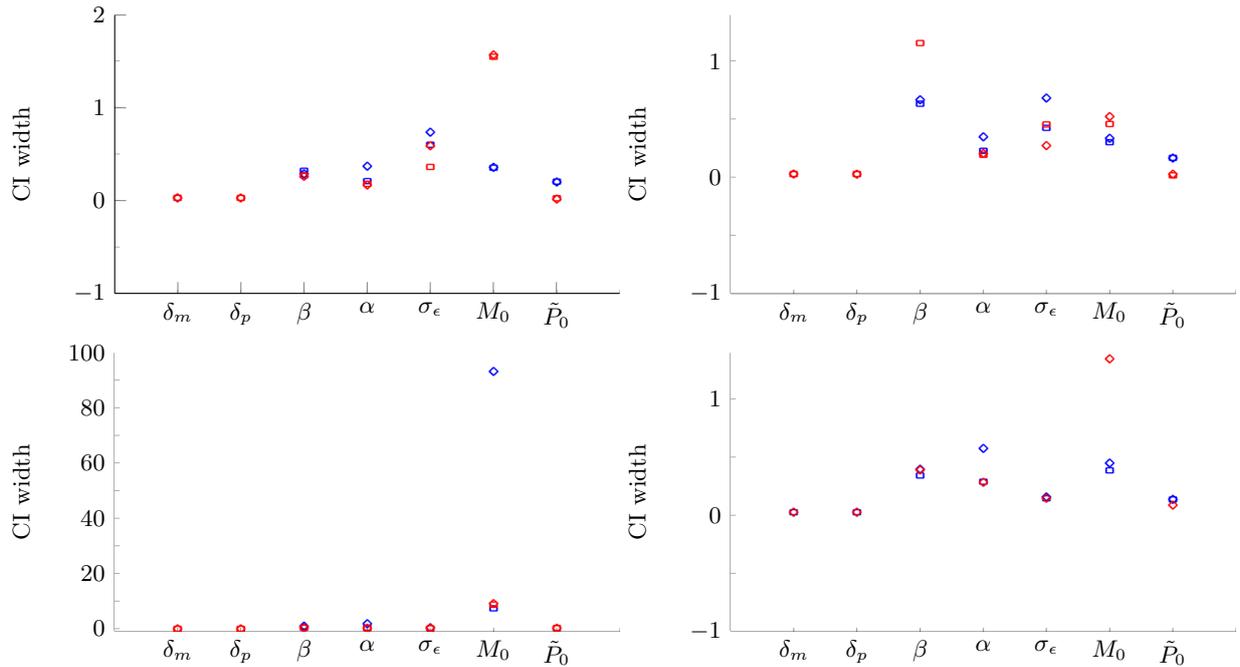


Figure 7: The width of the 50% credible interval calculated under the LNA (blue diamond), the LNA with κ fixed at the truth (blue square), the BDA with κ fixed at the truth (red diamond) and the BDA with κ fixed at the posterior median of the LNA. Each panel corresponds to a different set of parameters for systems of different molecular numbers as stated in the caption of Figure 6.

data, significantly more iterations were required to fully explore the BDA posterior (4.5 million iterations for the immature and 2.5 million iterations for the mature dataset) compared to the posterior under the LNA (200K and 400K, respectively) which, despite the cost of the LNA, corresponds to a much slower approach under the BDA. As discussed previously, obtaining informative priors for the degradation parameters is essential for model identifiability and can be obtained from additional experiments. In particular, we use the distributions obtained in Finkenstädt et al. (2013).

The resulting transcriptional profiles obtained under each of the approximations for both datasets are given in Figure 8. Comparing between the two datasets, we see that both tissues exhibit dynamic switching behaviour with multiple switching events occurring throughout the time course. The mature tissue shows greater variability in the transcription profiles and moreover transcription occurs at a higher rate than in the immature tissue. One hypothesis for this higher rate is that once removed from the animal, the mature tissue is released from dopamine inhibition which stimulates Prolactin production. This hypothesis also explains the higher synchronicity observed in the mature tissue compared to the immature tissue.

Figure 9 shows an estimated transcription profile for a single cell time course from the mature tissue under both the LNA and BDA along with the 95% credible intervals of the posterior switch times (vertical lines) and transcription rates (horizontal lines). This example typifies the differences observed through the LNA and BDA methods. Firstly, the estimated times of

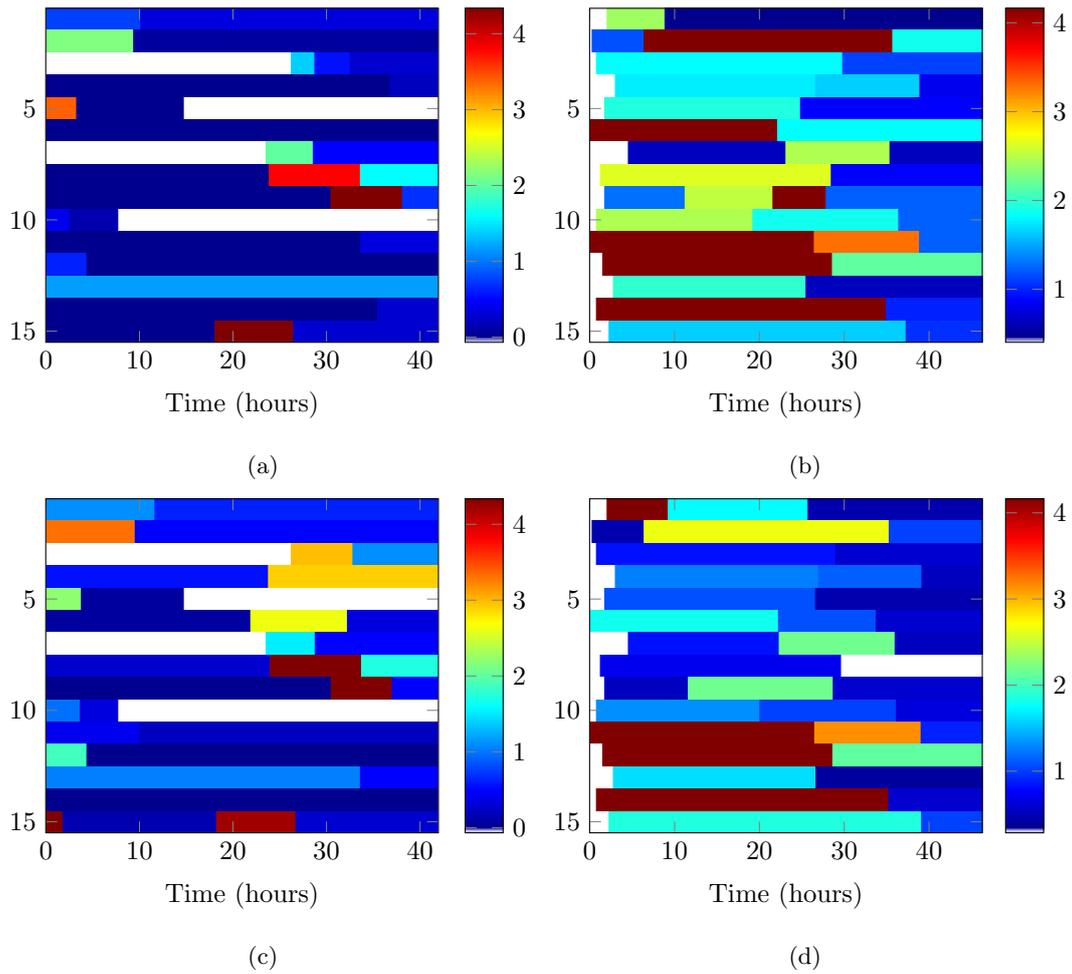
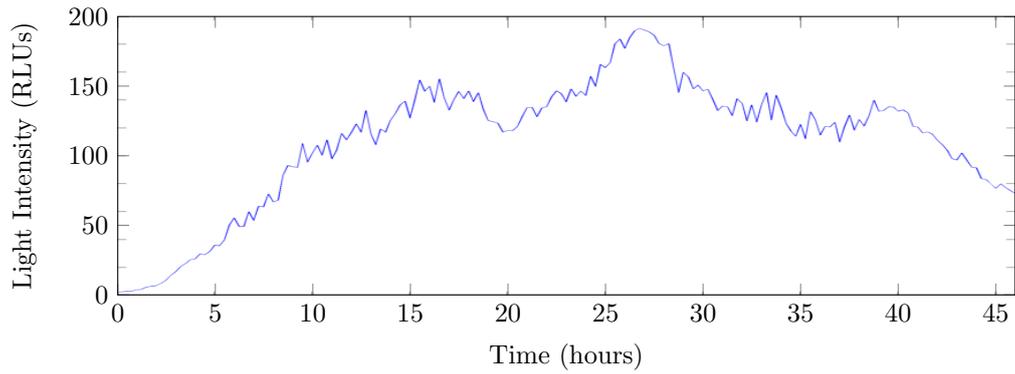
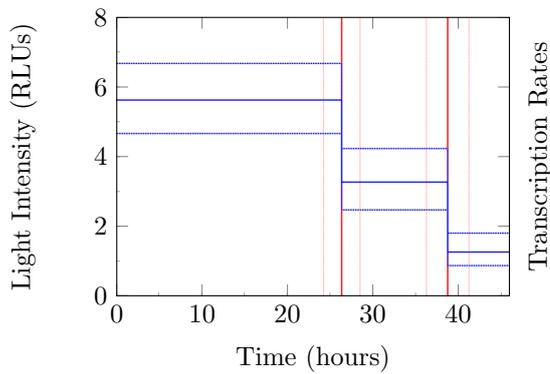


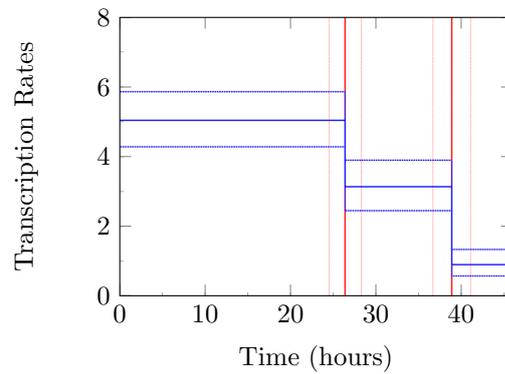
Figure 8: Heatmaps of the posterior transcriptional profiles for a) immature tissue sample and b) mature tissue sample calculated under the LNA. c) and d) are calculated under the BDA.



(a)



(b)



(c)

Figure 9: a) the raw time course data for a single cell from the mature tissue sample, with the backcalculated transcriptional profile given in b) under the LNA and in c) under the BDA. Dashed lines represent the 95% credible intervals about the posterior medians.

transcriptional events are highly consistent between the two approaches although there is an increased precision under the BDA. In addition, the 95% credible intervals for transcriptional rates are wider under the LNA as found in the simulation study.

In order to assess how well each of the approximations fit the data, we look at the following recursive residuals,

$$r_t = \frac{y_t - \mathbb{E}(Y_t|y_{1:t-1})}{\sqrt{\text{Var}(Y_t|y_{1:t-1})}}, \quad \text{for } t = 1, \dots, T, \quad (21)$$

where $Y_t|y^{1:t-1}$ is the one-step ahead predictive distribution. The latter is computed as part of the Kalman filter for the LNA, while under the BDA the moments of the predictive density can be represented by the weighted sample

$$\mathbb{E}(h(X_t)|y_{1:t-1}) = \frac{\sum_{i=1}^N w_i h(x_t^{(i)})}{\sum_{i=1}^N w_i}, \quad (22)$$

for weights w_1, \dots, w_N and samples $x_t^{(1)}, \dots, x_t^{(N)}$ and for any function h . Therefore, it is straightforward to extract the recursive residuals for both models. Under a state space model formulation the residuals in (21) will be i.i.d. with mean zero and variance one if the model fits the data. Moreover, if the state space formulation is Gaussian, the residuals will also be Gaussian. Figure 10 shows the residuals of the LNA and BDA models applied to the time series in Figure 9. The residuals were computed at the posterior median of all parameter values and we see that in both cases, they satisfy all assumptions, although there appears to be a slight increasing trend in the variance of the residuals under the LNA.

7 Discussion

In this study, we have proposed a general methodology for inferring transcriptional regulation for data obtained through single cell imaging techniques. The underlying biological model is flexible enough to describe a wide range of behaviours that provide insight into the underlying mechanisms. Moreover, a reversible jump MCMC scheme correctly captures the switching behaviour as shown through our simulation study. In order to achieve the above, we consider two approximations to the true stochastic system. Although less precise, the LNA has the advantage both in terms of the computational speed, but also in its ability to identify the scaling parameter of the measurement process. This parameter is of interest as it allows one to obtain an estimate of the underlying system size. However, since the BDA gives a more accurate representation of the stochastic system, we propose the use of this in conjunction with the LNA estimate of the scaling parameter. The BDA, although more expensive than the LNA, is still massively cheaper than the exact methods reviewed within this paper. This is because we continue to work with the underlying transition densities albeit through a normal approximation. Despite this, the BDA provides a realistic alternative to both the LNA and exact approaches when inferring systems of

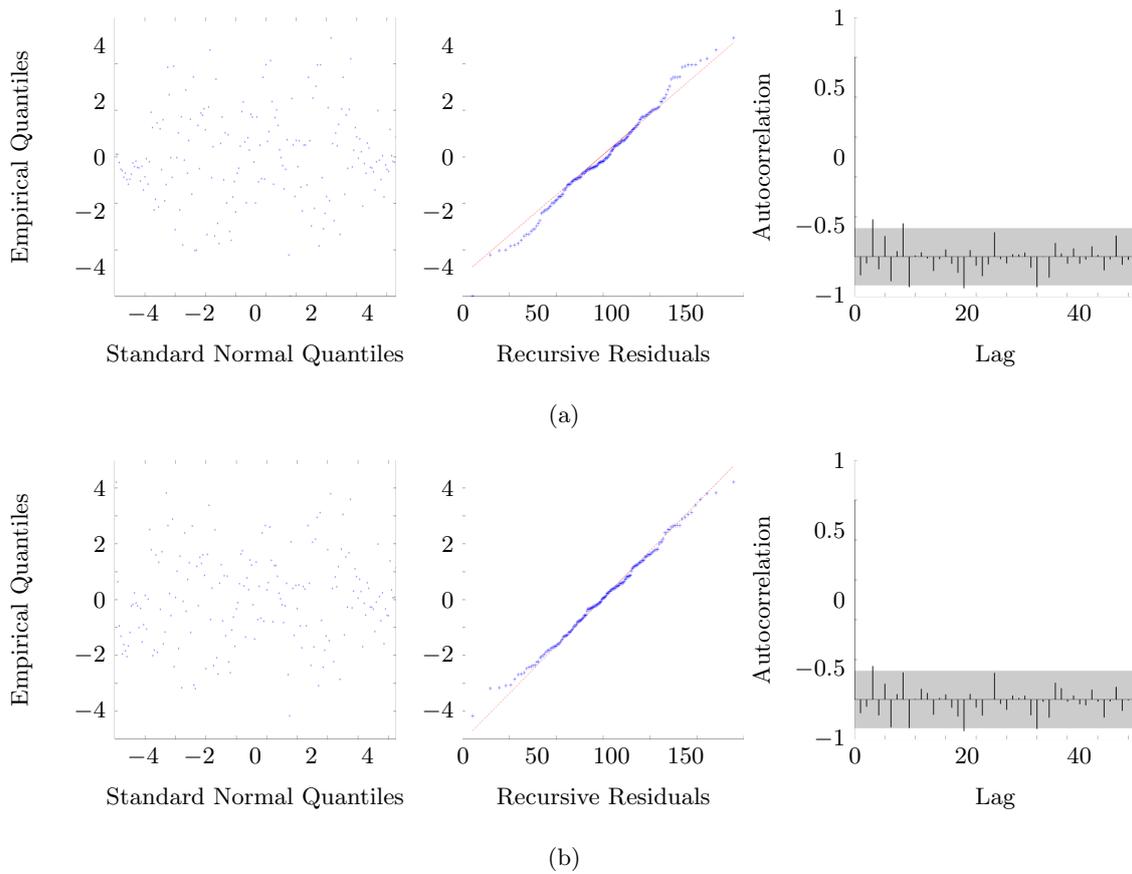


Figure 10: Recursive residuals calculated for the single time series shown in Figure 9 from the mature dataset calculated at the posterior median estimates obtained under a) the LNA and b) the BDA.

small molecular numbers. The BDA is specific to our gene expression model, however, many different stochastic reaction networks can be expressed as a sequence of conditionally independent birth-death reactions and a similar approach may be more widely applied.

Acknowledgments

K. Hey is supported by the EPSRC [ASTAA1112.KXH], H. Momiji is supported by the Wellcome Trust Grant [RSMAA.3020.SRA], K. Featherstone, J. Davis and M. White are supported by a Wellcome Trust Grant [67252], D. Rand is supported by an EPSRC Senior Fellowship [EP/C544587/1] and B. Finkenstadt is supported by BBSRC and EPSRC [GR/S29256/01, BB/F005814/1].

The authors wish to acknowledge helpful discussions with D. Jenkins and A. Finke on the statistical modelling and methodology and also A. Patist for discussions regarding the application.

Conflict of Interest: None declared.

References

- Amrein, M. and H. R. Künsch (2012). Rate estimation in partially observed Markov jump processes with measurement errors. *Statistics and Computing* 22(2), 513–526.
- Anderson, D. F. and T. G. Kurtz (2011). Continuous time Markov chain models for chemical reaction networks. In *Design and Analysis of Biomolecular Circuits*, pp. 3–42. Springer.
- Andrieu, C., A. Doucet, and R. Holenstein (2009). Particle Markov chain Monte Carlo for efficient numerical simulation. In *Monte Carlo and quasi-Monte Carlo methods 2008*, pp. 45–60. Springer.
- Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(3), 269–342.
- Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association* 95(452), 1269–1276.
- Boys, R. J., D. J. Wilkinson, and T. B. L. Kirkwood (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing* 18(2), 125–135.
- Chesson, P. (1978). Predator-prey theory and variability. *Annual Review of Ecology and Systematics* 9, 323–347.
- Daigle, B. J., M. K. Roh, L. R. Petzold, and J. Niemi (2012). Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics* 13(1), 68.

- Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain (2002). Stochastic gene expression in a single cell. *Science Signaling* 297(5584), 1183.
- Fearnhead, P. (2011). MCMC for State-Space Models. In Brooks, Gelman, Jones, and Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, pp. 513–529. Chapman & Hall/CRC Handbook of Modern Statistical Methods.
- Fearnhead, P., V. Giagos, and C. Sherlock (2014). Inference for reaction networks using the linear noise approximation. *Biometrics*.
- Finkenstädt, B., E. A. Heron, M. Komorowski, K. Edwards, S. Tang, C. V. Harper, J. R. E. Davis, M. R. H. White, A. J. Millar, and D. A. Rand (2008). Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics* 24(24), 2901–2907.
- Finkenstädt, B., D. J. Woodcock, M. Komorowski, C. V. Harper, J. R. E. Davis, M. R. H. White, and D. A. Rand (2013). Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: An application to single cell data. *The Annals of Applied Statistics* 7(4), 1960–1982.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. CRC press.
- Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* 81(25), 2340–2361.
- Gillespie, D. T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics* 113, 297.
- Golightly, A., D. A. Henderson, and C. Sherlock (2014). Delayed acceptance particle mcmc for exact inference in stochastic kinetic models. *arXiv preprint arXiv:1401.4369*.
- Golightly, A. and D. J. Wilkinson (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* 61(3), 781–788.
- Golightly, A. and D. J. Wilkinson (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* 1(6), 807–820.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82(4), 711–732.
- Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli* 7(2), 223–242.
- Harper, C. V., K. Featherstone, S. Semprini, S. Friedrichsen, J. McNeilly, P. Paszek, D. G. Spiller, A. S. McNeilly, J. J. Mullins, and J. R. E. Davis (2010). Dynamic organisation of prolactin gene expression in living pituitary tissue. *Journal Of Cell Science* 123(3), 424.

- Harper, C. V., B. Finkenstädt, D. J. Woodcock, S. Friedrichsen, S. Semprini, L. Ashall, D. G. Spiller, J. J. Mullins, D. A. Rand, and J. R. E. Davis (2011). Dynamic analysis of stochastic transcription cycles. *PLoS Biology* 9(4), e1000607.
- Heron, E. A., B. Finkenstädt, and D. A. Rand (2007). Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics* 23(19), 2596.
- Jenkins, D. J., B. Finkenstädt, and D. A. Rand (2013). A temporal switch model for estimating transcriptional activity in gene expression. *Bioinformatics* 29(9), 1158–1165.
- Komorowski, M., B. Finkenstädt, C. Harper, and D. Rand (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* 10(1), 343.
- Kurtz, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability* 8(2), 344–356.
- Kurtz, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Their Applications* 6(3), 223–240.
- Larson, D. R., R. H. Singer, and D. Zenklusen (2009). A single molecule view of gene expression. *Trends In Cell Biology* 19(11), 630–637.
- Opper, M. and G. Sanguinetti (2010). Learning combinatorial transcriptional dynamics from gene expression data. *Bioinformatics* 26(13), 1623–1629.
- Paulsson, J. (2004). Summing up the noise in gene networks. *Nature* 427, 415–418.
- Paulsson, J. (2005). Models of stochastic gene expression. *Physics of life reviews* 2(2), 157–175.
- Peccoud, J. and B. Ycart (1995). Markovian modeling of gene-product synthesis. *Theoretical Population Biology* 48(2), 222–234.
- Raj, A. and A. Van Oudenaarden (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135(2), 216–226.
- Sanchez, A., S. Choubey, and J. Kondev (2013). Stochastic models of transcription: from single molecules to single cells. *Methods* 62(1), 13–25.
- Sanguinetti, G., A. Rutter, M. Opper, and C. Archambeau (2009). Switching regulatory models of cellular stress response. *Bioinformatics* 25(10), 1280–1286.
- Stathopoulos, V. and M. A. Girolami (2013). Markov chain Monte Carlo inference for Markov jump processes via the linear noise approximation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371(1984).
- Stephens, D. J. and V. J. Allan (2003). Light microscopy techniques for live cell imaging. *Science Signaling* 300(5616), 82.

Suter, D. M., N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science* 332(6028), 472.

The MathWorks Inc. (2012). *MATLAB[®] version 7.12.0 (R2012a)*.

van Kampen, N. G. (1961). A power series expansion of the master equation. *Canadian Journal of Physics* 39(4), 551–567.

Wallace, E. W. J., D. T. Gillespie, K. R. Sanft, and L. R. Petzold (2012). Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *Systems Biology, IET* 6(4), 102–115.