BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology

Darren J. Wilkinson
Newcastle University, UK
d.j.wilkinson@ncl.ac.uk

Summary

Bacteria are single-celled organisms which often display heterogeneous behaviour, even among populations of genetically identical cells in uniform environmental conditions. Markov process models arising from the theory of stochastic chemical kinetics are often used to understand the genetic regulation of the behaviour of individual bacterial cells. However, such models often contain uncertain parameters which need to be estimated from experimental data. Parameter estimation for complex high-dimensional Markov process models using diverse, partial, noisy and poorly calibrated time-course experimental data is a challenging inferential problem, but a computationally intensive Bayesian approach turns out to be effective. The utility and added-value of the approach is demonstrated in the context of a stochastic model of a key cellular decision made by the gram-positive bacterium *Bacillus subtilis*, using quantitative data from single-cell fluorescence microscopy and flow cytometry experiments.

Keywords and Phrases: Bacillus subtilis; Genetic Regulation; GFP; Likelihood-free MCMC; Motility; Time-lapse Fluorescence Microscopy.

1. INTRODUCTION

Bacteria are single-celled prokaryotic organisms. Despite being relatively simple organisms, they often display complex heterogeneous behaviour, even among populations of genetically identical cells in uniform environmental conditions (Wilkinson 2009). Markov process models arising from the theory of stochastic chemical kinetics (Wilkinson 2006) are often used to understand the genetic regulation of the behaviour of individual bacterial cells. However, such models often contain uncertain parameters which need to be estimated from experimental data. Parameter estimation for complex high-dimensional Markov process models using diverse, partial, noisy and poorly calibrated time-course experimental data is a challenging

inferential problem, but several previous studies have demonstrated that progress is possible (Golightly & Wilkinson 2005, 2006, Boys et al. 2008, Henderson et al. 2009). It will be demonstrated here that a computationally intensive Bayesian approach can, in principle, be effective for understanding the information in the data regarding plausible parameter values. The utility and added-value of the approach will demonstrated in the context of a stochastic model of a key cellular decision, the decision to become motile, made by the gram-positive bacterium *Bacillus subtilis*. The inferential issues will be illustrated using simulated data based on single-cell fluorescence microscopy and flow cytometry experiments.

2. BACTERIAL GENE REGULATION

2.1. Bacillus subtilis

Bacillus subtilis (Sonenshein et al. 2002) is the most widely studied model gram positive bacterium. It is relatively easy to culture in the lab, and is highly genetically tractable, being naturally competent for genetic transformation (Dubnau 1991). It was the first gram positive bacterium to be sequenced, and its genome is relatively well characterised (Moszer et al. 2002). *B. subtilis* has a relatively interesting life cycle, and must make expensive cellular decisions on the basis of the information it has regarding its environment. The default behaviour for a *B. subtilis* cell in a rich nutrient environment is to grow and divide, but in response to certain stresses it may choose to become competent for genetic transformation (Dubnau 1991), sporulate (Errington 1993), or become motile (Kearns & Losick 2005).

2.2. *Motility regulation*

One of the key decisions a *B. subtilis* cell must make is whether or not to grow flagella and become motile (Kearns & Losick 2005), leading to the possibility of swimming away from its current location to a new and better environment. Like most other decision systems in living organisms, the precise details of how this decision is made is extremely complex. In this paper we will focus on one small aspect of this problem, in order to illustrate the important concepts without getting lost in biological complexity.

Bacteria typically use special proteins called σ factors in order to regulate transcription. Most genes cannot be transcribed (are turned off) unless an appropriate σ factor is available. The *B. subtilis* sigma factor σ^D is key for the regulation of motility. Many of the genes and operons encoding motility-related proteins are governed by this σ factor, and so understanding its regulation is key to understanding the motility decision. The gene for σ^D is embedded in a large operon containing several other motility-related genes, known as the *fla/che* operon. The *fla/che* operon itself is under the control of another σ factor, σ^A, but is also regulated by other proteins. In particular, transcription of the operon is strongly repressed by the protein *CodY*, which is encoded upstream of *fla/che*. *CodY* inhibits transcription by binding to the *fla/che* promoter. Since *CodY* is upregulated in good nutrient conditions, this is thought to be a key mechanism for motility regulation.

As previously mentioned, many motility-related genes are under the control of σ^D. For simplicity we focus here on one such gene, *hag*, which encodes the protein *flagellin* (or *Hag*), the key building block of the flagella. It so happens that *hag* is also directly repressed by *CodY*. The regulation structure can be illustrated using the simple schematic given in Figure 1. It should be emphasised that this is only one small component of the regulation of motility, and that a great deal more is

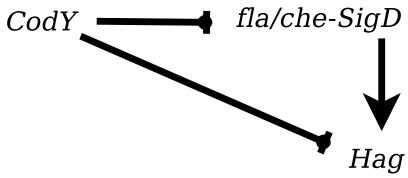

Figure 1: A small component of the regulation of motility in *B. subtilis*

known about the complex regulation of motility than is presented here. However, the aspect presented here is sufficient to illustrate the essential statistical issues.

3. MODELLING AND INFERENCE

3.1. *Stochastic kinetic models*

Computational systems biology (Kitano 2002) is concerned with developing dynamic simulation models of biological processes such as the motility regulation network model previously described. Such models are useful for developing a quantitative understanding of the process, for testing current understanding of the mechanisms, and to allow *in silico* experimentation that would be difficult or time consuming to carry out on the real system in the lab. Traditionally, continuous deterministic models were developed, typically using an assumption of mass-action chemical kinetics leading to systems of ordinary differential equations. However, in recent years there has been increasing recognition of the importance of modelling intrinsic stochasticity in intra-cellular biological processes, not captured by the traditional approaches (Wilkinson 2009). The theory of stochastic chemical kinetics forms the basis of a more realistic class of models, which models cellular dynamics using a Markov jump process (Wilkinson 2006).

For mass-action stochastic kinetic models, it is assumed that the state of the system at a given time is represented by the number of molecules of each reacting chemical "species" present in the system at that time, and that the state of the system is changed at discrete times according to one or more reaction "channels". We assume there are u species denoted $\mathcal{X}_1, \ldots, \mathcal{X}_u$, and v reactions, $\mathcal{R}_1, \ldots, \mathcal{R}_v$. Each reaction $\mathcal{R}_i$ is of the form

$$p_{i1}\mathcal{X}_1 + \cdots + p_{iu}\mathcal{X}_u \longrightarrow q_{i1}\mathcal{X}_1 + \cdots + q_{iu}\mathcal{X}_u, \ i = 1, \ldots, v.$$

Here p_{ij} denotes the number of molecules of $\mathcal{X}_j$ that will be *consumed* by reaction $\mathcal{R}_i$, and q_{ij} the number of molecules *produced*. Let P be the $v \times u$ matrix formed from the p_{ij} and Q be the corresponding matrix of the q_{ij}. We can write the entire

reaction system in matrix/vector form as

$$P\mathcal{X} \longrightarrow Q\mathcal{X}.$$

The matrices P and Q are typically *sparse*, and this fact can be exploited in computational algorithms. The $u \times v$ matrix $S = (Q - P)^\mathsf{T}$ is the *stoichiometry matrix* of the system, and is especially important in computational analysis of stochastic kinetic models, as its columns encode the change of state in the system caused by the different reaction events. Let X_{jt} denote the number of molecules of $\mathcal{X}_j$ at time t and $X_t = (X_{1t}, \ldots, X_{ut})^\mathsf{T}$. We assume that reaction $\mathcal{R}_i$ has *hazard* (or *rate law*, or *propensity*) $h_i(X_t, c_i)$, where c_i is a *rate parameter*. We put $c = (c_1, \ldots, c_v)^\mathsf{T}$ and $h(X_t, c) = (h_1(X_t, c_1), \ldots, h_v(X_t, c_v))^\mathsf{T}$ in order to simplify notation. Under certain assumptions (Gillespie 1992), it can be shown that the system evolves as a *Markov jump process* with independent reaction hazards for each reaction channel. Further, for *mass-action stochastic kinetics*, the algebraic form of each rate laws is given as

$$h_i(X_t, c_i) = c_i \prod_{j=1}^{u} \binom{X_{jt}}{p_{ij}}, \quad i = 1, \ldots, v.$$

Hence, given a reaction network structure, the vector of reaction rate constants, c, determines the stochastic behaviour of the system.

A mathematical representation of this Markov jump process can be constructed, known as the *random time change representation* (Kurtz 1972), which turns out to be very helpful for mathematical analysis of the system. Let R_{it} denote the number reactions of type $\mathcal{R}_i$ in the time window $(0, t]$, and then define $R_t = (R_{1t}, \ldots, R_{vt})^\mathsf{T}$. It should be clear that $X_t - X_0 = SR_t$ (this is known as the *state updating equation*). Now for $i = 1, \ldots, v$, define $N_i(t)$ to be the count functions for v independent *unit Poisson processes*. Then

$$R_{it} = N_i\left(\int_0^t h_i(X_\tau, c_i)d\tau\right).$$

Putting $N(t_1, \ldots, t_v) = (N_1(t_1), \ldots, N_v(t_v))^\mathsf{T}$, we can write

$$R_t = N\left(\int_0^t h(X_\tau, c)d\tau\right)$$

to get

$$X_t - X_0 = S\,N\left(\int_0^t h(X_\tau, c)d\tau\right),$$

the random time-change representation of the Markov jump process. See Ball et al. (2006) for applications of this representation to analysis of approximate system dynamics.

This process is typically non-linear with unbounded state space. Consequently the models are generally analytically intractable, but realisations of the model can be simulated exactly using a computer, using a discrete event simulation algorithm, known in this context as the Gillespie algorithm (Gillespie 1977). The inference, or inverse problem, is to determine plausible values for the rate constants, c, from partial, discrete and noisy observations of the system state.

3.2. *Bayesian inference for complex Markov process models*

3.2.1. *Concepts and notation*

At some level, inference for complex Markov process models is not fundamentally more difficult than for many other high-dimensional non-linear statistical models. Given complete information about the trajectory of the process over a given fixed time window, the likelihood of the process can be computed exactly. If we observe the process $\boldsymbol{x} = \{x(t) : t \in [0,T]\}$ where $x(t)$ represents the values of X_t for one particular (observed) realisation of the stochastic process, we can determine from the reaction structure the time (t_i) and type (ν_i) of the n reaction events occurring in the time interval $(0,T]$. Suppose that the ith reaction event is (t_i, ν_i), $i = 1, \ldots n$. Also define $t_0 = 0, t_{n+1} = T$. Let r_j be the total number of type j events occurring (so $n = \sum_{j=1}^{v} r_j$). Then the complete-data likelihood for the observed sample path is

$$L(c;\boldsymbol{x}) \equiv \Pr(\boldsymbol{x} \mid c) = \left\{ \prod_{i=1}^{n} h_{\nu_i}(x(t_{i-1}), c_{\nu_i}) \right\} \exp\left\{ -\int_0^T h_0(x(t), c)\, dt \right\}.$$

See Chapter 10 of Wilkinson (2006) for further details. Note that the integral occurring in the above equation is just a finite sum, so there are no computational issues associated with evaluating it (though as usual, it is numerically advantageous to actually work with the log of the likelihood).

There are further simplifications which arise for rate laws of the form $h_i(x, c_i) = c_i g_i(x)$ (true for basic mass-action stochastic kinetic models), as then the complete-data likelihood factorises as

$$L(c;\boldsymbol{x}) = \prod_{j=1}^{v} L_j(c_j;\boldsymbol{x})$$

where

$$L_j(c_j;\boldsymbol{x}) = c_j^{r_j} \exp\left\{ -c_j \int_0^T g_j(x(t))\, dt \right\}, \quad j = 1, \ldots, v.$$

These component likelihoods are semi-conjugate to priors of the form $c_j \sim \Gamma(a_j, b_j)$ and hence can be combined to get full-conditional posterior distributions of the form

$$c_j \mid \boldsymbol{x} \sim \Gamma\left(a_j + r_j, b_j + \int_0^T g_j(x(t)) dt \right).$$

All of the inferential complications arise from the fact that, in practise, we cannot hope to observe the system perfectly over any finite time window. Observations of the system state will typically occur at discrete times, will usually be partial (not all species in the model will be measured), and will often be subject to measurement error. This data-poor scenario leads to a challenging missing-data problem. Consider first the best-case scenario — perfect observation of the system at discrete times. Conditional on discrete-time observations, the Markov process breaks up into a collection of independent bridge processes that appear not to be analytically tractable. We can attempt to use MCMC to explore sample paths consistent with the end-points of the random intervals. Considering just one interval, we need to explore r_t consistent with $x_{t+1} - x_t = Sr_t$. Both reversible jump and

block-updating strategies are possible — see Boys et al. (2008) for details, but these standard MCMC techniques do not scale well to large, complex models with very large numbers of reaction events.

One way forward is to approximate the true Markov jump process by a diffusion process, known in this context as the *chemical Langevin equation* (CLE) (Gillespie 2000). Then techniques for Bayesian estimation of stochastic differential equation models can be applied (Golightly & Wilkinson 2005, 2006, 2008), but this approach too is far from straightforward, and for many interesting problems the diffusion approximation will be unsatisfactory.

3.2.2. *Likelihood-free MCMC*

One of the problems with the above approaches to inference in realistic data-poor scenarios is the difficulty of developing algorithms to explore a huge (discrete) state space with a complex likelihood structure that makes conditional simulation difficult. Such problems arise frequently, and in recent years interest has increasingly turned to methods which avoid some of the complexity of the problem by exploiting the fact that we are easily able to forward-simulate realisations of the process of interest. Methods such as likelihood-free MCMC (LF-MCMC) (Marjoram et al. 2003) and Approximate Bayesian Computation (ABC) (Beaumont et al. 2002) are now commonly used to tackle problems which would be extremely difficult to solve otherwise.

A likelihood-free approach to this problem can be constructed as follows. Let $\pi(\mathbf{x} \,|\, c)$ denote the (complex) likelihood of the simulation model. Let $\pi(\mathcal{D} \,|\, \mathbf{x}, \tau)$ denote the (simple) measurement error model, giving the probability of observing the data $\mathcal{D}$ given the output of the stochastic process and some additional parameters, τ. Put $\theta = (c, \tau)$, and let $\pi(\theta)$ be the prior for the model parameters. Then the joint density can be written

$$\pi(\theta, \mathbf{x}, \mathcal{D}) = \pi(\theta)\pi(\mathbf{x} \,|\, \theta)\pi(\mathcal{D} \,|\, \mathbf{x}, \theta).$$

Suppose that interest lies in the posterior distribution $\pi(\theta, \mathbf{x} \,|\, \mathcal{D})$. A Metropolis-Hastings scheme can be constructed by proposing a joint update for θ and $\mathbf{x}$ as follows. Supposing that the current state of the Markov chain is $(\theta, \mathbf{x})$, first sample a proposed new value for θ, $\theta^\star$, by sampling from some (essentially) arbitrary proposal distribution $f(\theta^\star \,|\, \theta)$. Then, *conditional on this newly proposed value*, sample a proposed new sample path, $\mathbf{x}^\star$ by forwards simulation from the model $\pi(\mathbf{x}^\star \,|\, \theta^\star)$. Together the newly proposed pair $(\theta^\star, \mathbf{x}^\star)$ is accepted with probability $\min\{1, A\}$, where

$$A = \frac{\pi(\theta^\star)}{\pi(\theta)} \times \frac{f(\theta \,|\, \theta^\star)}{f(\theta^\star \,|\, \theta)} \times \frac{\pi(\mathcal{D} \,|\, \mathbf{x}^\star, \theta^\star)}{\pi(\mathcal{D} \,|\, \mathbf{x}, \theta)}.$$

Crucially, the potentially problematic likelihood term, $\pi(\mathbf{x} \,|\, \theta)$ does not occur in the acceptance probability, due to the fact that a sample from it was used in the construction of the proposal. Note that choosing an independence proposal of the form $f(\theta^\star \,|\, \theta) = \pi(\theta^\star)$ leads to the simpler acceptance ratio

$$A = \frac{\pi(\mathcal{D} \,|\, \mathbf{x}^\star, \theta^\star)}{\pi(\mathcal{D} \,|\, \mathbf{x}, \theta)}.$$

This "canonical" choice of proposal also lends itself to more elaborate schemes, as we will consider shortly.

This "vanilla" LF-MCMC scheme should perform reasonably well provided that $\mathcal{D}$ is not high-dimensional, and there is sufficient "noise" in the measurement process to make the probability of acceptance non-negligible. However, in practice $\mathcal{D}$ is often of sufficiently large dimension that the overall acceptance rate of the scheme is intolerably low. In this case it is natural to try and "bridge" between the prior and the posterior with a sequence of intermediate distributions. There are several ways to do this, but here it is most natural to exploit the Markovian nature of the process and consider the sequence of posterior distributions obtained as each additional time point is observed. For notational simplicity consider equispaced observations at integer times and define the data up to time t as $\mathcal{D}_t = \{d_1, \ldots, d_t\}$. Similarly, define sample paths $\mathbf{x}_t \equiv \{x_s \,|\, t-1 < s \leq t\}$, $t = 1, 2, \ldots$, so that $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$. The posterior at time t can then be computed inductively as follows.

(i) Assume at time t we have a (large) sample from $\pi(\theta, x_t \,|\, \mathcal{D}_t)$ (for time 0, initialise with sample from prior)

(ii) Run an MCMC algorithm which constructs a proposal in two stages:

  (a) First sample $(\theta^\star, x_t^\star) \sim \pi(\theta, x_t \,|\, \mathcal{D}_t)$ by picking at random and perturbing $\theta^\star$ slightly (sampling from a kernel density estimate of the distribution)

  (b) Next sample $\mathbf{x}_{t+1}^\star$ by forward simulation from $\pi(\mathbf{x}_{t+1}^\star \,|\, \theta^\star, x_t^\star)$

  (c) Accept/reject $(\theta^\star, x_{t+1}^\star)$ with probability $\min\{1, A\}$ where

$$A = \frac{\pi(d_{t+1} \,|\, x_{t+1}^\star, \theta^\star)}{\pi(d_{t+1} \,|\, x_{t+1}, \theta)}$$

(iii) Output the sample from $\pi(\theta, x_{t+1} \,|\, \mathcal{D}_{t+1})$, put $t := t+1$, return to step 2.

Consequently, for each observation d_t, an MCMC algorithm is run which takes as input the current posterior distribution prior to observation of d_t and outputs the posterior distribution given all observations up to d_t. As d_t is typically low-dimensional, this strategy usually leads to good acceptance rates.

It is worth emphasising the generality of this algorithm. Although we are here applying it to stochastic kinetic models, it is applicable to any Markov process discretely observed with error. It is also trivially adaptable to non-uniform observations, and to observation of multiple independent time courses (the posterior distribution from one time course can be used to form the prior distribution for the next). It is also adaptable to data from multiple models which share many parameters — an important scenario in systems biology, as we shall see later.

3.2.3. *CaliBayes*

The sequential likelihood-free algorithm described above can be implemented in a reasonably generic manner. The resulting algorithms are very powerful, but exceptionally computationally intensive. It is therefore natural to want exploit powerful remote computing resources connected to a local machine via the Internet. CaliBayes (`http://www.calibayes.ncl.ac.uk/`) is an example of such a remote facility. Simulation models (either deterministic or stochastic) are encoded using the Systems Biology Markup Language (SBML) (Hucka et al. 2003), and these are sent to the remote server together with a large sample from the prior distribution and

the experimental data. When the computations are completed, a large sample from the posterior distribution is returned to the user. The CaliBayes system uses a service-oriented architecture (SOA), and makes use of modern web-service technology — further details are provided in Chen et al. (2010). The forward simulation of SBML models is carried out using third-party simulators such as COPASI (Hoops et al. 2006), FERN (Erhard et al. 2008) or BASIS (Kirkwood et al. 2003), and these may be specified by the user. An R package (`calibayesR`) which provides a user-friendly interface to most of the CaliBayes services is available from R-forge (`http://r-forge.r-project.org/`).

3.2.4. *Approximate Bayesian computation*

There is a close connection between LF-MCMC methods and those of approximate Bayesian computation (ABC). Consider first the case of a perfectly observed system, so that there is no measurement error model. Then there are model parameters θ described by a prior $\pi(\theta)$, and a forwards-simulation model for the data $\mathcal{D}$, defined by $\pi(\mathcal{D} \,|\, \theta)$. It is clear that a simple algorithm for simulating from the desired posterior $\pi(\theta \,|\, \mathcal{D})$ can be obtained as follows. First simulate from the joint distribution $\pi(\theta, \mathcal{D})$ by simulating $\theta^\star \sim \pi(\theta)$ and then $\mathcal{D}^\star \sim \pi(\mathcal{D} \,|\, \theta^\star)$. This gives a sample $(\theta^\star, \mathcal{D}^\star)$ from the joint distribution. A simple rejection algorithm which rejects the proposed pair unless $\mathcal{D}^\star$ matches the true data $\mathcal{D}$ clearly gives a sample from the required posterior distribution. However, in many problems this will lead to an intolerably high rejection rate. The "approximation" is to accept values provided that $\mathcal{D}^\star$ is "sufficiently close" to $\mathcal{D}$. In the simplest case, this is done by forming a (vector of) summary statistic(s), $s(\mathcal{D}^\star)$ (ideally a *sufficient statistic*), and accepting provided that $|s(\mathcal{D}^\star) - s(\mathcal{D})| < \varepsilon$ for some suitable choice of metric and ε (Beaumont et al. 2002). However, in certain circumstances this "tolerance", ε can be interpreted as a measurement error model (Wilkinson 2008), and for problems involving large amount of data, ABC may be applied sequentially (Sisson et al. 2007). Sequential ABC approaches have been applied to systems biology problems by Toni et al. (2009). Further, it is well known that ABC approaches can be combined with MCMC to get approximate LF-MCMC schemes (Marjoram et al. 2003).

4. MOTILITY REGULATION MODEL

4.1. *Model structure*

The essential relationships central to the model for motility regulation depicted in Figure 1 can be translated into a set of biochemical reactions as given in Table 1. The usual convention of starting names of genes with lower case letters and the corresponding proteins with upper case letters has been adopted. Again note that for illustrative purposes, many simplifications have been made in this model. In particular, the processes of transcription, translation, folding and protein maturation have been collapsed into a single reaction step.

Given specification of the initial conditions of the system and all reaction rate constants, it is straightforward to simulate realisations from the associated Markov jump process model using the Gillespie algorithm. A typical trajectory starting from zero protein molecules is given in Figure 2. We can use simulated trajectories of this nature in order to understand the associated inferential problem. Again, to keep the problem as simple as possible, we will assume that just three rate constants are uncertain, and that these are the object of inference, using appropriate time course

Table 1: Basic reaction structure for the motility regulation model

$$
\begin{aligned}
\mathsf{codY} &\longrightarrow \mathsf{codY} + \mathsf{CodY} \\
\mathsf{CodY} &\longrightarrow \emptyset \\
\mathsf{flache} &\longrightarrow \mathsf{flache} + \mathsf{SigD} \\
\mathsf{SigD} &\longrightarrow \emptyset \\
\mathsf{SigD\_hag} &\longrightarrow \mathsf{SigD} + \mathsf{hag} + \mathsf{Hag} \\
\mathsf{Hag} &\longrightarrow \emptyset \\
\mathsf{SigD} + \mathsf{hag} &\longrightarrow \mathsf{SigD\_hag} \\
\mathsf{SigD\_hag} &\longrightarrow \mathsf{SigD} + \mathsf{hag} \\
\mathsf{CodY} + \mathsf{flache} &\longrightarrow \mathsf{CodY\_flache} \\
\mathsf{CodY\_flache} &\longrightarrow \mathsf{CodY} + \mathsf{flache} \\
\mathsf{CodY} + \mathsf{hag} &\longrightarrow \mathsf{CodY\_hag} \\
\mathsf{CodY\_hag} &\longrightarrow \mathsf{CodY} + \mathsf{hag}
\end{aligned}
$$

data. The three "unknowns" and their corresponding true values are

$$kSigDprod = 1,\ kflacherep = 0.02,\ kflacheunrep = 0.1.$$

They correspond to the maximal rate of production of SigD, and the binding and unbinding of CodY to the *fla/che* operon, respectively. These are plausibly the parameters of greatest scientific interest in the context of this model. The specification of sensible prior distributions for rate constants is a non-trivial problem (Liebermeister & Klipp 2005), but here we will adopt independent finite uniform priors on the log scale, as these have proven to be useful in applied work (Henderson et al. 2010):

$$
\begin{aligned}
\log(kSigDprod) &\sim \mathrm{Unif}(\log\{0.01\}, \log\{100\}), \\
\log(kflacherep) &\sim \mathrm{Unif}(\log\{0.0002\}, \log\{2\}), \\
\log(kflacheunrep) &\sim \mathrm{Unif}(\log\{0.001\}, \log\{10\}).
\end{aligned}
$$

These priors cover two orders of magnitude either side of the true value, and hence represent very vague prior knowledge.

4.2. *Single-cell time course data*

4.2.1. *Observation of* σ^D

We will start by assuming that it is possible to directly observe the number of molecules of σ^D in a single cell over time. Observations will be made every 5

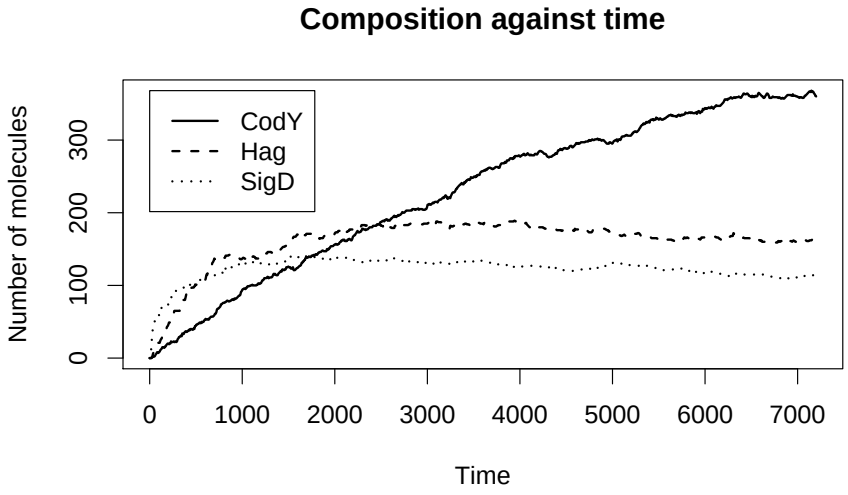

Figure 2: A typical realisation of the motility model

minutes (300 seconds) for 2 hours (7,200 seconds) giving a total of 24 observations. We make the simplifying (and unrealistic) assumption that the initial state of the cell is known. We assume that the measurements are subject to a small amount of measurement error that is I.I.D. Gaussian with a known standard deviation of 10 molecules.

It is straightforward to apply the LF-MCMC algorithm described in Section 3.2.2 to this problem. Here, 1,000,000 particles were used, together with a burn-in of 1,000 iterations and a thin of 5, so that, in total, 5,001,000 MCMC iterations are performed per observation. These figures were sufficient to give adequate coverage and low autocorrelations in the particle chain.

The marginal posterior distributions for the three parameters of interest are shown in Figure 3 (top). The [5%, 50%, 95%] quantiles of the marginals for $kSigDprod$, $kflacherep$ and $kflacheunrep$ are $[-0.13,\ 0.90,\ 2.66]$, $[-5.93,\ -1.97,\ 0.45]$ and $[-4.86,\ -1.72,\ 1.07]$, respectively. It is clear that there is a great deal of information in the data regarding the likely value of $kSigDprod$ — the maximum rate of production of σ^D, but apparently much less about the other two parameters. This is somewhat misleading, as the two parameters are partially confounded and have high posterior correlation as shown in Figure 3 (middle). The data therefore clearly contains a reasonable amount of information about all three parameters. Figure 3 (bottom) shows in grey 90% equitailed pointwise posterior predictive probability intervals for the key model species, with the (unknown, unobserved) true values overlaid. Clearly the interval for σ^D is tight around the true values, as this is the observed species, but the other two species are also reasonably well identified by the observed data (and the model). Note that if further information is required, pooling observations from multiple cells is straightforward, as the parameter posterior from

one cell can be used as the parameter prior for the next in a natural sequential manner.

4.2.2. *Observation of* Hag

It turns out not to be completely straightforward to observe levels of σ^D directly, partly because the σ^D gene is embedded in the middle of the large *fla/che* operon. Before examining in detail exactly how measurements are typically made, it is instructive to consider observation of *Hag*, which has its own promoter, and is strongly activated by σ^D. We consider the same observation protocol as above, but this time use (noisy) measurements of *Hag* levels in order to make inferences about the three key unknowns.

The marginal posterior distributions for the three parameters of interest given data on *Hag* are shown in Figure 4 (top). The [5%, 50%, 95%] quantiles of the marginals are [0.29, 1.76, 3.61], [−6.32, −2.26, 0.41] and [−6.58, −4.01, −0.32], respectively. These inferences are broadly consistent with the inference obtained by observing σ^D, but there is less information in the *Hag* data than in the corresponding data for σ^D.

4.2.3. *Time-lapse microscopy and* GFP *reporters*

In fact, it turns out not to be straightforward to accurately measure *any* native protein directly. To observe and track gene expression in single living cells over time, some kind of *reporter system* is typically employed. Although there are alternatives, fluorescent reporters are often used, with *green fluorescent protein* (*GFP*) being the most common. *GFP* was originally isolated from a jellyfish, and can be detected in single living cells with a fluorescence camera attached to a powerful microscope if the cells are first exposed to UV light. The gene for *GFP*, *gfp*, has to be integrated into the host genome in such a way as to try to make the levels of mature *GFP* correlate strongly with the levels of the target protein of interest. This often turns out to be technically difficult, and less-than-perfect alternatives are often employed.

In the case of σ^D, the standard strategy is to form a fusion of the promoter of *hag*, P_{hag} to *gfp*, to get P_{hag}-*gfp*, and then integrate this construct into a convenient place in the genome, which is often at the locus known as *amyE*. The genotype of the resulting mutant is typically written *amyE*::P_{hag}-*gfp* (Kearns & Losick 2005). The rationale behind this construction is that P_{hag} is strongly activated by σ^D, and so when levels of σ^D are high, the production rate of *GFP* should also be high. Note however, that there is absolutely no reason to suppose a linear relationship between the levels of σ^D and the level of *GFP*, and hence the measured levels of fluorescence. There are several additional sources of discrepancy, including the fact that *GFP* is a relatively stable protein, and therefore decays more slowly than most other proteins. Additionally, since the *amyE* locus is close to the origin of replication, there will typically be two copies of this gene per cell, whereas the *hag* and σ^D genes are far from the origin, and hence will typically be single-copy only. Although there clearly is a relationship between the levels of σ^D and *GFP*, this relationship must be explicitly modelled in a quantitative way. Some actual time lapse microscopy images of cells of this genotype are shown in Figure 5. Images such as these must be analysed to track individual cells over time, and to quantify the levels of *GFP* fluorescence in each cell at each time point. Specialist image analysis algorithms (Wang et al. 2010) can be used to automate this process.

The additional species and reactions can be added into the model considered previously, and the SBML-shorthand (Wilkinson 2006) corresponding to the full

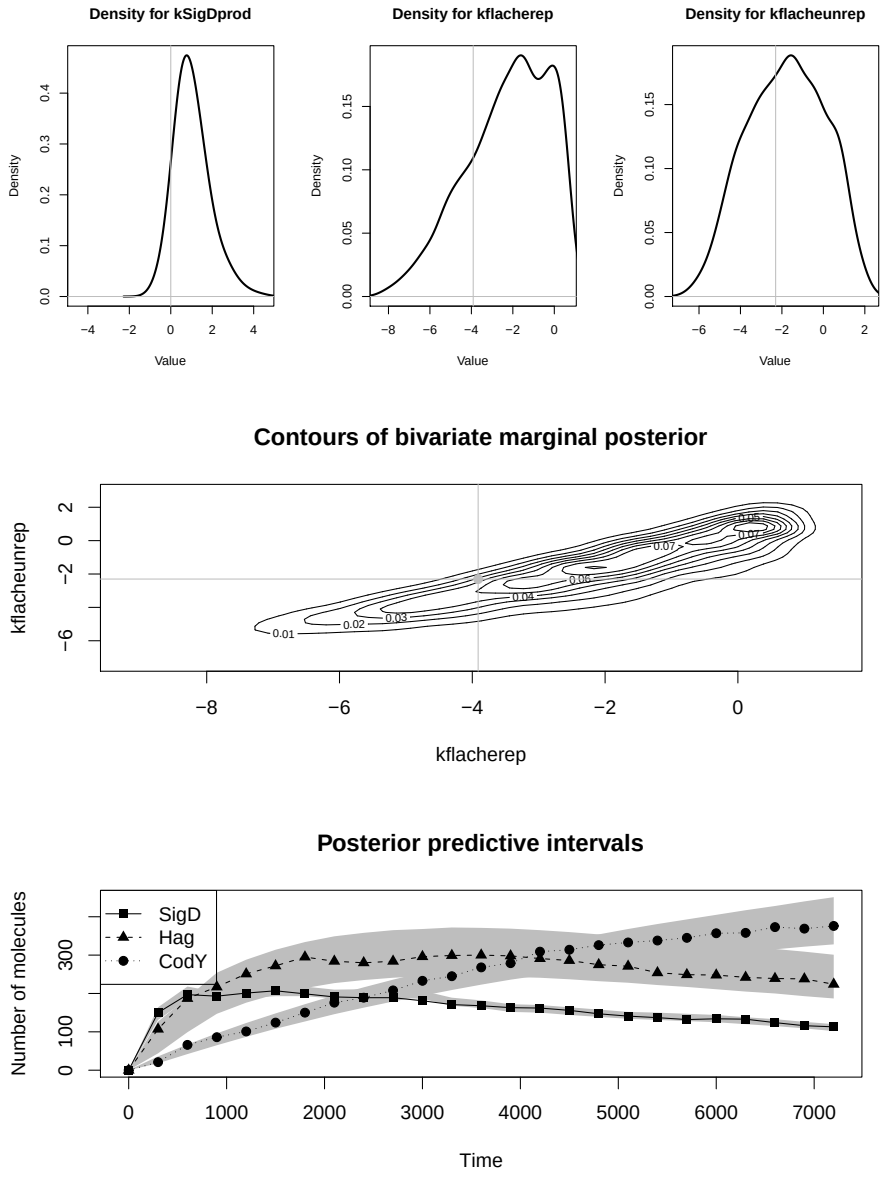

Figure 3: Top: Marginal posterior distributions for the (log of the) three parameters of interest, based on 24 observations of σ^D. True value shown as a vertical line. Middle: Contour plot of the bivariate posterior distribution of the (log of the) *fla/che* binding and unbinding constants. True value shown as the intersection of the two lines. Bottom: Predictive distributions for the key model species (90%, equitailed, pointwise) in grey, with true (unknown) values overlaid.

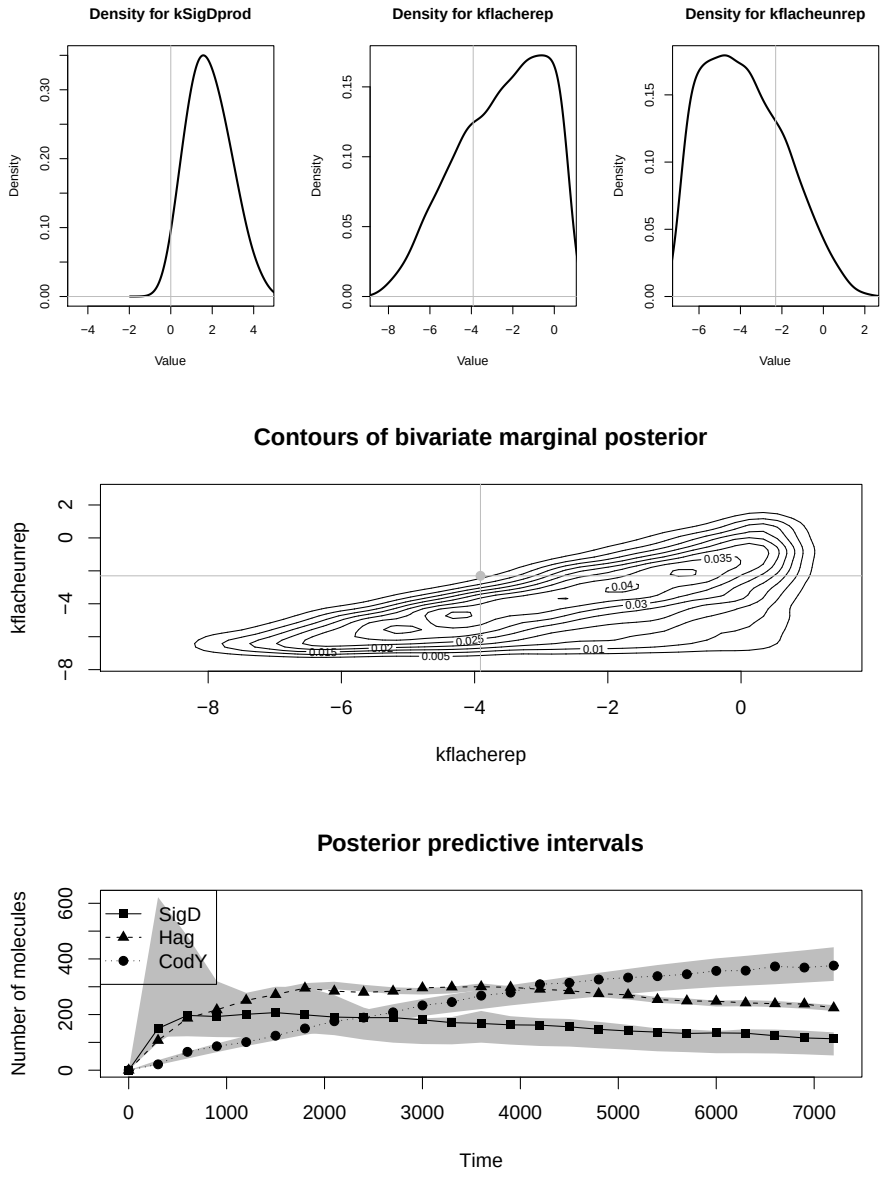

Figure 4: Top: Marginal posterior distributions for the (log of the) three parameters of interest, based on 24 observations of *Hag*. Middle: Contour plot of the bivariate posterior distribution of the (log of the) *fla/che* binding and unbinding constants. Bottom: Predictive distributions for the key model species, with true (unknown) values overlaid.

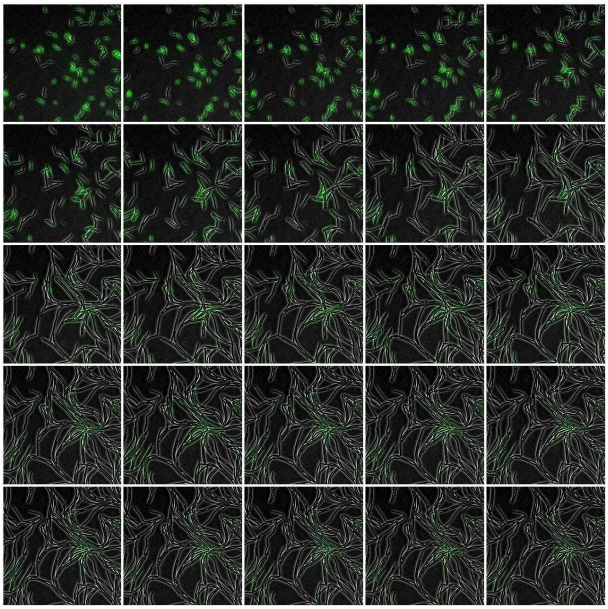

Figure 5: Time-lapse microscopy images of growing and dividing *B. subtilis* cells with genotype *amyE*::P_{hag}-*gfp*. Experiment conducted by the author using a DeltaVision microscopy system during a visit to the lab of Dr Leendert Hamoen (Newcastle).

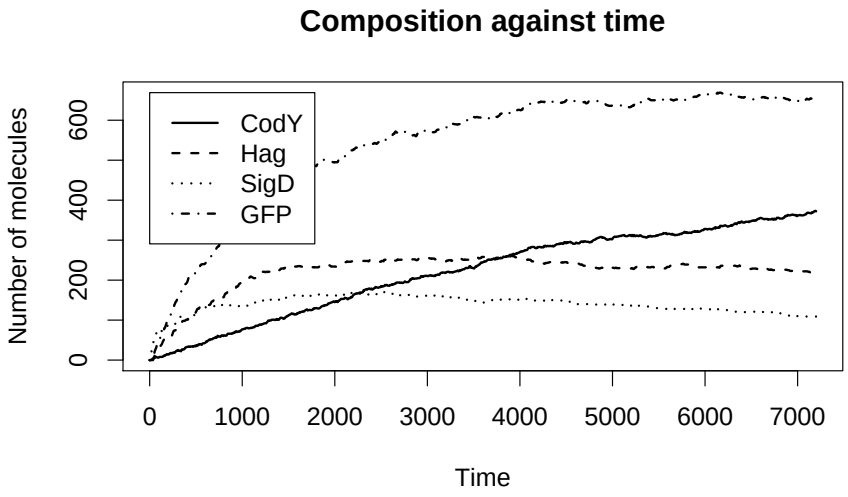

Figure 6: A typical realisation of the motility model, including the GFP reporter

resulting model is given in the Appendix. A typical realisation from this full model is shown in Figure 6, showing the relationship between a few of the key species. Note the less-than-perfect relationship between the levels of σ^D and GFP.

Inference for this enlarged model can be carried out using the same LF-MCMC algorithm as previously described. Again, assuming 24 measurements of GFP levels (subjecting the cells to UV light more than once every 5 minutes is toxic), inference for the three key unknowns can proceed as before.

The marginal posterior distributions obtained using this extended model for the three parameters of interest are shown in Figure 7 (top). The [5%, 50%, 95%] quantiles of the marginals are [0.14, 1.35, 3.31], [−6.91, −2.69, 0.37] and [−6.23, −3.14, 0.23], respectively.

Although the GFP data is not quite as informative about the model parameters as direct observations of levels of σ^D would be, considerable information can still be gained. See Finkenstadt et al. (2008) for related work based on a linear noise approximation. It is natural to wonder whether it is worth the effort of modelling GFP levels explicitly as we have done here, rather than simply assuming that the GFP levels correspond to levels of σ^D. We can examine this question by re-running our inferential procedure for measurements on σ^D, but using the actual measured levels of GFP.

The marginal posterior distributions for the three parameters of interest are shown in Figure 8 (top). The [5%, 50%, 95%] quantiles of the marginals are [−0.36, −0.08, 0.22], [−5.88, −3.62, −1.91] and [−1.81, 0.36, 2.11], respectively. This (incorrect) posterior distribution is potentially misleading. There appears to be very strong information regarding $kSigDprod$ — more information than we re-

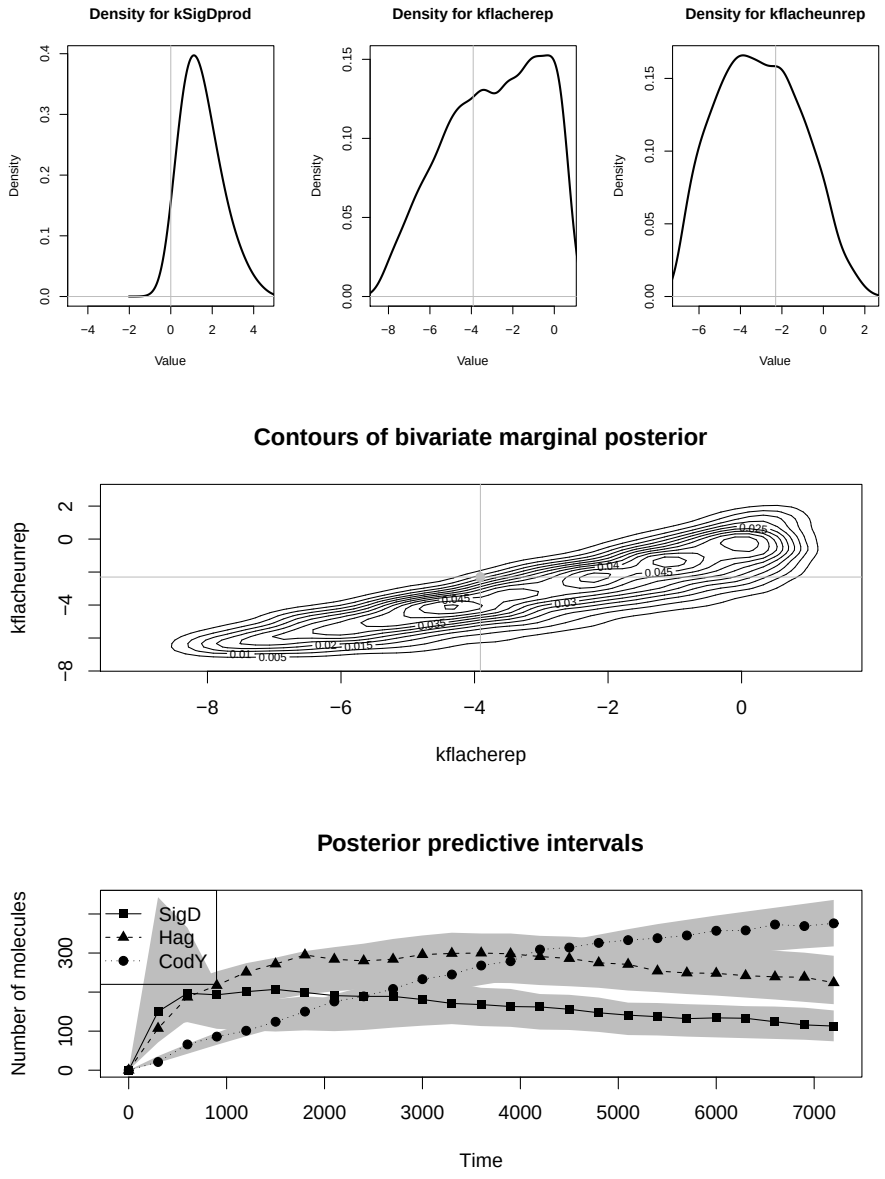

Figure 7: Top: Marginal posterior distributions for the (log of the) three parameters of interest, based on 24 observations of *GFP*. Middle: Contour plot of the bivariate posterior distribution of the (log of the) *fla/che* binding and unbinding constants. Bottom: Predictive distributions for the key model species, with true (unknown) values overlaid.

ally have. It so happens in this case that the posterior contains the true value, but that is simply a consequence of the fact that the rates of production of *GFP* and σ^D are assumed to be the same in this model. Further, the posteriors for the other two parameters are not correctly centred on the true parameter values — the true parameter values are very unlikely according to this posterior distribution. The filtering distributions are also (obviously) very badly calibrated. Thus, quantitative modelling of the relationship between measured GFP levels and the target protein of interest is clearly worthwhile.

There is a further potential complication with the use of fluorescence (and luminescence) data that has not yet been discussed. Although there is reason to believe that the measured fluorescence intensity will be in direct proportion to the number of molecules of mature *GFP*, often the data is *uncalibrated* in the sense that the constant of proportionality is (at least partially) unknown. Often it is possible to get a good handle on it using calibration data, but in general it will be desirable to include this constant as a further model parameter — see Henderson et al. (2010) for an example. Furthermore, it is not even completely clear that the measured fluorescence is in fact directly proportional to the number of *GFP* molecules, as there is some suggestion that at high concentration the *GFP* molecules form aggregates which are not fluorescent (Iafolla et al. 2008).

4.3. *Population data and knock-out variants*

Ultimately, obtaining just one read-out on one particular protein is inevitably going to be limited in terms of the information that can be obtained. There are several obvious strategies to improve this situation. The first is to use multiple reporters in the same cells. This can be accomplished by using different coloured fluorescent reporters for different proteins of interest. In principle it is possible to use up to around four such reporters within a cell using current technology, but in practice it seems to be technically difficult to use more than two reliably. Another useful technique is to obtain data from cells with key genes knocked out. Provided that the gene is non-essential, it is easy to construct the model corresponding to the knock-out, and this new model will have many parameters in common with the original. Data from multiple models can be combined sequentially by taking the posterior for relevant parameters from one model as priors for the next.

Time-lapse microscopy is currently the only practical way to track expression in individual cells over time. However, there are other technologies, such as *flow cytometry*, which can take measurements on thousands of individual cells at a given time. This technology can be used to monitor how the *distribution* of expression in a population changes over time (and in different knock-outs). This data too is informative for model parameters, and is an effective alternative to time-lapse microscopy in certain situations. There are several ways that such population level data can be used for model parameter inference. Perhaps the simplest (but computationally intensive) method is to use the ABC techniques described in Section 3.2.4 in conjunction with ensemble forward simulations from the model, conditioning by checking whether the simulated distribution of measurements is sufficiently close to the observed distribution, under some suitable metric on empirical distributions.

5. SUMMARY

This paper has shown how Markov process models can be used to understand the stochastic dynamics of bacterial gene regulation. Inference for model parameters

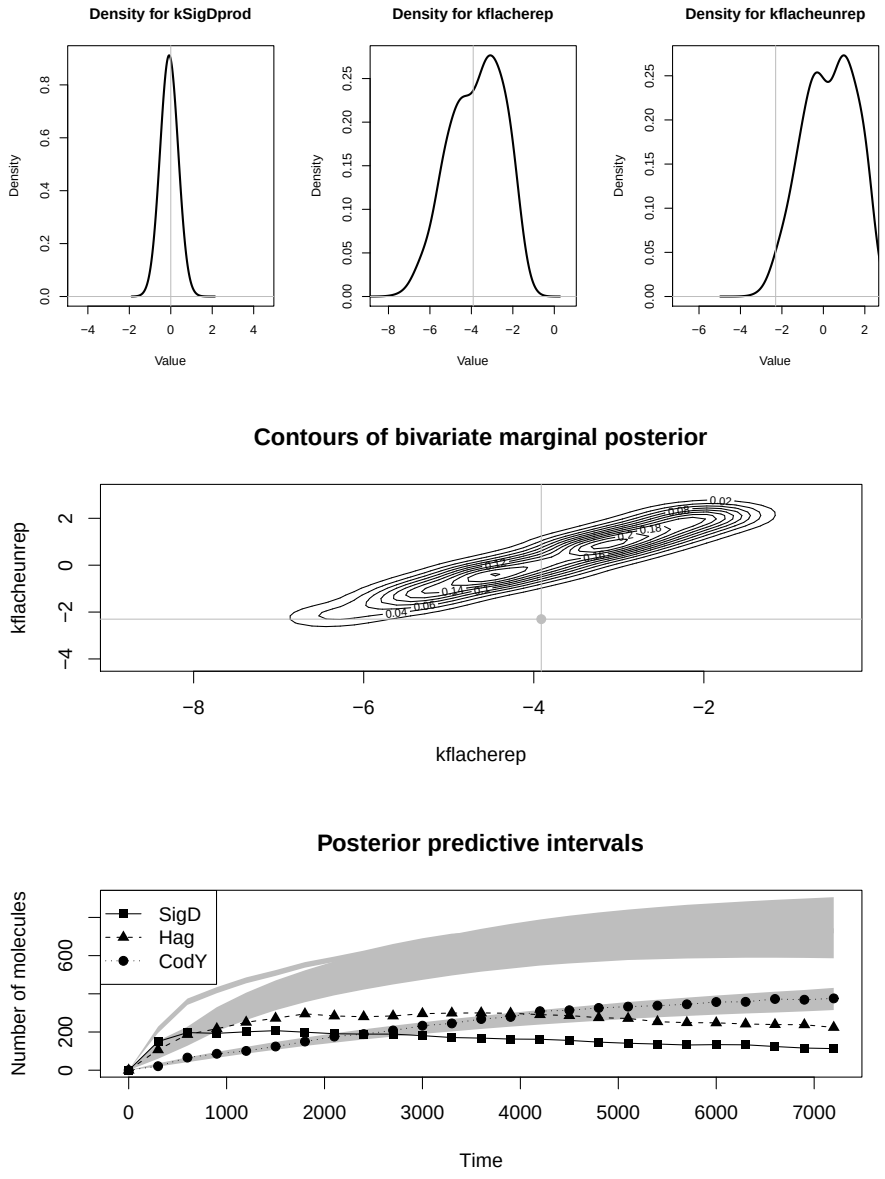

Figure 8: Top: Marginal posterior distributions for the (log of the) three parameters of interest, based on 24 observations of *GFP* treated (incorrectly) as observations of σ^D. Middle: Contour plot of the bivariate posterior distribution of the (log of the) *fla/che* binding and unbinding constants. Bottom: Predictive distributions for the key model species, with true (unknown) values overlaid.

from time-course measurements of system state is an important problem, and computationally intensive Bayesian algorithms such as LF-MCMC and ABC have been shown to be useful here due to their inherent flexibility. Explicit quantitative modelling of the measurement process (including the relationship between fluorescent reporters and their target proteins) has been shown to be an important and non-ignorable aspect of the modelling process. There is clearly still a long way to go before such techniques can be routinely used in practice as part of a systems biology approach. Combining time-lapse data from multiple experiments, mutants and conditions, together with similar data from flow cytometry experiments, for parameter estimation and model comparison, is still technically challenging, and the experimental systems themselves require improvement and calibration in order to be suitable for fully quantitative analysis. Integrating these single-cell analyses with other molecular biology technologies such as microarrays and RNA-Sequencing data is a further challenge. However, many of the issues to be faced are fundamentally statistical in nature, and so it seems that statisticians have an important role to play in advancing current biological knowledge.

ACKNOWLEDGEMENTS

This work was funded by the Biotechnology and Biological Sciences Research Council through grants BBF0235451, BBSB16550 and BBC0082001. The author would also like to thank Dr Leendert Hamoen and members of his lab for hosting a visit by the author during the first half of 2009.

REFERENCES

Ball, K., Kurtz, T. G., Popovic, L. & Rempala, G. (2006) , 'Asymptotic analysis of multiscale approximations to reaction networks', *Annals of Applied Probability* **16**(4), 1925–1961.

Beaumont, M. A., Zhang, W. & Balding, D. J. (2002) , 'Approximate Bayesian computation in population genetics', *Genetics* **162**(4), 2025–2035.

Boys, R. J., Wilkinson, D. J. & Kirkwood, T. B. L. (2008) , 'Bayesian inference for a discretely observed stochastic kinetic model', *Statistics and Computing* **18**(2), 125–135.

Chen, Y., Lawless, C., Gillespie, C. S., Wu, J., Boys, R. J. & Wilkinson, D. J. (2010) , 'CaliBayes and BASIS: integrated tools for the calibration, simulation and storage of biological simulation models', *Briefings in Bioinformatics* **11**(3), 278–289.

Dubnau, D. (1991) , 'Genetic competence in Bacillus subtilis.', *Microbiology and Molecular Biology Reviews* **55**, 395–424.

Erhard, F., Friedel, C. C. & Zimmer, R. (2008) , 'FERN - a Java framework for stochastic simulation and evaluation of reaction networks.', *BMC Bioinformatics* **9**, 356.

Errington, J. (1993) , 'Bacillus subtilis sporulation: regulation of gene expression and control of morphogenesis.', *Microbiology and Molecular Biology Reviews* **57**, 1–33.

Finkenstadt, B., Heron, E. A., Komorowski, M., Edwards, K., Tang, S., Harper, C. V., Davis, J. R., White, M. R., Millar, A. J. & Rand, D. A. (2008) , 'Reconstruction of transcriptional dynamics from gene reporter data using differential equations.', *Bioinformatics* **24**(24), 2901–2907.

Gillespie, D. T. (1977) , 'Exact stochastic simulation of coupled chemical reactions', *Journal of Physical Chemistry* **81**, 2340–2361.

Gillespie, D. T. (1992) , 'A rigorous derivation of the chemical master equation', *Physica A* **188**, 404–425.

Gillespie, D. T. (2000) , 'The chemical Langevin equation', *Journal of Chemical Physics* **113**(1), 297–306.

Golightly, A. & Wilkinson, D. J. (2005) , 'Bayesian inference for stochastic kinetic models using a diffusion approximation', *Biometrics* **61**(3), 781–788.

Golightly, A. & Wilkinson, D. J. (2006) , 'Bayesian sequential inference for stochastic kinetic biochemical network models', *Journal of Computational Biology* **13**(3), 838–851.

Golightly, A. & Wilkinson, D. J. (2008) , 'Bayesian inference for nonlinear multivariate diffusion models observed with error', *Computational Statistics and Data Analysis* **52**(3), 1674–1693.

Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C. & Wilkinson, D. J. (2009) , 'Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons', *Journal of the American Statistical Association* **104**(485), 76–87.

Henderson, D. A., Boys, R. J., Proctor, C. J. & Wilkinson, D. J. (2010) , Linking systems biology models to data: a stochastic kinetic model of p53 oscillations, *in* A. O'Hagan & M. West, eds, 'Handbook of Applied Bayesian Analysis', Oxford University Press.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P. & Kummer, U. (2006) , 'COPASI — a complex pathway simulator', *Bioinformatics* **22**(24), 3067–3074.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novere, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. & Wang, J. (2003) , 'The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models', *Bioinformatics* **19**(4), 524–531.

Iafolla, M. A., Mazumder, M., Sardana, V., Velauthapillai, T., Pannu, K. & McMillen, D. R. (2008) , 'Dark proteins: effect of inclusion body formation on quantification of protein expression.', *Proteins* **72**(4), 1233–1242.

Kearns, D. B. & Losick, R. (2005) , 'Cell population heterogeneity during growth of Bacillus subtilis', *Genes and Development* **19**, 3083–3094. 10.1101/gad.1373905.

Kirkwood, T. B. L., Boys, R. J., Gillespie, C. S., Proctor, C. J., Shanley, D. P. & Wilkinson, D. J. (2003) , 'Towards an e-biology of ageing: integrating theory and data', *Nature Reviews Molecular Cell Biology* **4**(3), 243–249.

Kitano, H. (2002) , 'Computational systems biology', *Nature* **420**(6912), 206–210.

Kurtz, T. G. (1972) , 'The relationship between stochastic and deterministic models for chemical reactions', *The Journal of Chemical Physics* **57**(7), 2976–2978.

Liebermeister, W. & Klipp, E. (2005) , 'Biochemical networks with uncertain parameters', *IEE Systems Biology* **152**(3), 97–107.

Marjoram, P., Molitor, J., Plagnol, V. & Tavare, S. (2003) , 'Markov chain Monte Carlo without likelihoods.', *Proc. Natl. Acad. Sci. U.S.A.* **100**(26), 15324–15328.

Moszer, I., Jones, L. M., Moreira, S., Fabry, C. & Danchin, A. (2002) , 'Subtilist: the reference database for the Bacillus subtilis genome', *Nucleic acids research* **30**, 62–65. 10.1093/nar/30.1.62.

Sisson, S. A., Fan, Y. & Tanaka, M. M. (2007) , 'Sequential Monte Carlo without likelihoods.', *Proc. Natl. Acad. Sci. U.S.A.* **104**(6), 1760–1765.

Sonenshein, A. L., Hoch, J. A. & Losick, R., eds (2002) , *Bacillus subtilis and its closest relatives*, ASM Press.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. (2009) , 'Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems', *J. R. Soc. Interface* **6**(31), 187–202.

Wang, Q., Niemi, J., Tan, C. M., You, L. & West, M. (2010) , 'Image segmentation and dynamic lineage analysis in single-cell fluorescence microscopy.', *Cytometry A* **77**(1), 101–110.

Wilkinson, D. J. (2006) , *Stochastic Modelling for Systems Biology*, Chapman & Hall/CRC Press, Boca Raton, Florida.

Wilkinson, D. J. (2009) , 'Stochastic modelling for quantitative description of heterogeneous biological systems', *Nature Reviews Genetics* **10**, 122–133. 10.1038/nrg2509.

Wilkinson, R. D. (2008) , Approximate Bayesian computation (ABC) gives exact results under the assumption of model error, Technical report, Sheffield University.

APPENDIX

MOTILITY MODEL

The full SBML-shorthand (Wilkinson 2006) for the model considered in this paper is given below. This can be converted to full SBML (Hucka at al. 2003) using the tools available from:

`http://www.staff.ncl.ac.uk/d.j.wilkinson/software/sbml-sh/`

```
@model:2.1.1=BSMod02 "Bacillus subtilis motility with GFP"
@units
 substance=item
@compartments
 Cell=1
@species
 Cell:codY=1 s
 Cell:CodY=0 s
 Cell:flache=1 s
 Cell:SigD=0 s
 Cell:hag=1 s
 Cell:Hag=0 s
 Cell:CodY_flache=0 s
 Cell:CodY_hag=0 s
 Cell:SigD_hag=0 s
 Cell:Phag_gfp=2 s
 Cell:SigD_Phag_gfp=0 s
 Cell:CodY_Phag_gfp=0 s
 Cell:GFP=0 s
@parameters
 kProtDeg=0.0002
 kCodOn=0.02
 kCodOff=0.1
 kProdSigD=1
@reactions
@r=CodYprod
 codY->codY+CodY
 k*codY : k=0.1
@r=CodYdeg
 CodY->
 kProtDeg*CodY
@r=SigDprod
 flache->flache+SigD
 kProdSigD*flache
@r=SigDdeg
 SigD->
 kProtDeg*SigD
@r=Hagprod
 SigD_hag->SigD+hag+Hag
 k*SigD_hag : k=1
```

```
@r=Hagdeg
 Hag->
 kProtDeg*Hag
@r=hagact
 SigD+hag->SigD_hag
 k*SigD*hag : k=0.01
@r=haginact
 SigD_hag->SigD+hag
 k*SigD_hag : k=0.1
@r=flacherep
 CodY+flache->CodY_flache
 kCodOn*CodY*flache
@r=flacheunrep
 CodY_flache->CodY+flache
 kCodOff*CodY_flache
@r=hagrep
 CodY+hag->CodY_hag
 k*CodY*hag : k=0.01
@r=hagunrep
 CodY_hag->CodY+hag
 k*CodY_hag : k=0.1
@r=GFPprod
 SigD_Phag_gfp->SigD+Phag_gfp+GFP
 k*SigD_Phag_gfp : k=1
@r=GFPdeg
 GFP->
 0.5*kProtDeg*GFP
@r=Phag_gfpact
 SigD+Phag_gfp->SigD_Phag_gfp
 k*SigD*Phag_gfp : k=0.01
@r=Phag_gfpinact
 SigD_Phag_gfp->SigD+Phag_gfp
 k*SigD_Phag_gfp : k=0.1
@r=Phag_gfprep
 CodY+Phag_gfp->CodY_Phag_gfp
 k*CodY*Phag_gfp : k=0.01
@r=Phag_gfpunrep
 CodY_Phag_gfp->CodY+Phag_gfp
 k*CodY_Phag_gfp : k=0.1
```