

Comparison of Parameter Estimation Methods in Stochastic Chemical Kinetic Models: Examples in Systems Biology

Ankur Gupta and James B. Rawlings

Dept. of Chemical and Biological Engineering, University of Wisconsin–Madison, 1415 Engineering Drive, Madison, WI 53705

DOI 10.1002/aic.14409

Published online March 5, 2014 in Wiley Online Library (wileyonlinelibrary.com)

Stochastic chemical kinetics has become a staple for mechanistically modeling various phenomena in systems biology. These models, even more so than their deterministic counterparts, pose a challenging problem in the estimation of kinetic parameters from experimental data. As a result of the inherent randomness involved in stochastic chemical kinetic models, the estimation methods tend to be statistical in nature. Three classes of estimation methods are implemented and compared in this paper. The first is the exact method, which uses the continuous-time Markov chain representation of stochastic chemical kinetics and is tractable only for a very restricted class of problems. The next class of methods is based on Markov chain Monte Carlo (MCMC) techniques. The third method, termed conditional density importance sampling (CDIS), is a new method introduced in this paper. The use of these methods is demonstrated on two examples taken from systems biology, one of which is a new model of single-cell viral infection. The applicability, strengths and weaknesses of the three classes of estimation methods are discussed. Using simulated data for the two examples, some guidelines are provided on experimental design to obtain more information from a limited number of measurements. © 2014 American Institute of Chemical Engineers *AICHE J*, 60: 1253–1268, 2014

Keywords: parameter estimation, systems biology, stochastic chemical kinetics, computational modeling, Bayesian inference

Introduction

The classical method to simulate the time evolution of reacting molecules is based on the continuum assumption. When the number of molecules of all reacting species in a set of chemical reactions is of the order of Avogadro's number, the concentration can be assumed to be a continuous real variable. In such cases, classic mass action kinetics can be used to describe the rates of reaction. When the number of molecules of one or more species is on the order of hundreds or thousands, however, we can no longer use the continuum assumption. As a result, instead of real-valued concentrations we need to consider the integer-valued number of molecules. Another effect of such low number of molecules is that the classical mass action kinetics is no longer valid. The reaction rates are no longer deterministic, and a probabilistic approach is required. Instead of accounting for amount of reactants consumed (and products produced) in a time interval, we need to account for the probability that a reaction occurs in a time interval. This approach of modeling chemical reactions has come to be known as stochastic chemical kinetics.^{1–4}

Another difference between the deterministic and stochastic regimes of chemical kinetics is seen in the simulation methods. While the classical, deterministic regime involves solving a system of coupled ordinary differential equations (ODEs), the stochastic regime requires a probabilistic method involving repeated generation of random numbers. Various algorithms have been developed to simulate stochastic chemical kinetics starting with the stochastic simulation algorithm (SSA), also known as Gillespie's algorithm.^{2,5} Note that unlike the solution of a system of ODEs, the SSA produces a different (random) trajectory for every simulation. Considerable research has been devoted to developing alternative algorithms^{6,7} and approximations with applications to various problems.^{8–12}

Cellular processes in systems biology frequently demonstrate intrinsic noise or randomness,^{13–15} which is caused by the presence of low number of reacting molecules. Arkin and coworkers^{16,17} used stochastic chemical kinetic models to explain how a homogeneous population of λ -infected *E. coli* cells can partition into subpopulations that follow different pathways. Weinberger et al.¹⁸ demonstrated that the otherwise unexplained phenotypic bifurcation observed in HIV-1 infected T-cells could be accounted for by the presence of low molecular concentrations and a stochastic transcriptional feedback loop. Hensel et al.¹⁹ developed a detailed model of intracellular growth of vesicular stomatitis virus (VSV) to demonstrate that stochastic gene expression contributes to the variation in viral yields.

The acceptance of stochastic reaction kinetics as a standard approach to explain intrinsic noise and the advances in

Additional Supporting Information may be found in the online version of this article.

Correspondence concerning this article should be addressed to James B. Rawlings at rawlings@engr.wisc.edu.

© 2014 American Institute of Chemical Engineers

experimental measurement techniques, especially fluorescence microscopy, led to the inverse problem of estimating parameters from data. Golding et al.²⁰ used fluorescence microscopy to measure the number of mRNA molecules in a single cell at a sampling rate of two measurements per minute. They showed that the average number of mRNA transcripts per cell was between zero and 10 and proposed a stochastic gene expression model to explain the observed behavior.

Initial attempts to estimate model parameters (usually the reaction rate constants) were rather *ad hoc*—the parameters were set to biologically plausible values and then tuned by eye so that the model simulation resembled the experimental data.¹⁷ Another approach is to use the continuum assumption and treat the reaction kinetics as deterministic instead of stochastic. In this case, parameter estimates can be easily obtained by least-squares fitting or maximum likelihood estimation. Tian et al.²¹ demonstrated that this approach does not produce good parameter estimates when the number of molecules is small. Later approaches maintained the stochasticity of the model and utilized various statistical Monte Carlo methods to estimate parameters. These estimation methods may be classified into two familiar categories—maximum likelihood methods and Bayesian inference methods. Maximum likelihood methods include simulated maximum likelihood,²¹ density function distance,²² and approximate maximum likelihood^{22,23} methods. All of these methods maximize an approximation of the true likelihood to obtain parameter estimates. Stochastic gradient descent, proposed by Wang et al.,²⁴ estimates the gradients of the likelihood function with respect to the parameters using a reversible jump Markov chain Monte Carlo (RJMCMC) method. By contrast, Bayesian inference methods attempt to obtain the posterior distribution of the parameters, which can then be maximized to obtain maximum *a posteriori* (MAP) estimates. Most Bayesian inference methods rely on MCMC techniques. Rempala et al.²⁵ utilize MCMC-enabled Bayesian inference to estimate parameters in a specific model of gene transcription. Wilkinson and coworkers^{26–30} have developed a range of Bayesian inference algorithms, all of which use MCMC methods. In many cases, they replace the stochastic chemical kinetic model with another statistical model²⁶ or approximate it using the diffusion approximation.^{27–29} In other cases, they use the true stochastic model.^{30,31} Similar to the MCMC method of Wilkinson,³⁰ Choi and Rempala³² proposed a Bayesian inference algorithm based on Gibbs sampling. All of these estimation methods suffer from their own particular shortcomings and have restricted applicability. We will discuss these issues in this article.

Many of the examples in the references above come from systems biology—Prokaryotic gene-autoregulation,^{17,24,29}

gene transcription models,^{23,25} RNA expression,^{20,22,23} and single-cell mtDNA dynamics.²⁶ Other examples include the Lotka–Volterra model^{29–31} and epidemic models.³² These models consist of 2–8 reactions involving 2–5 species.

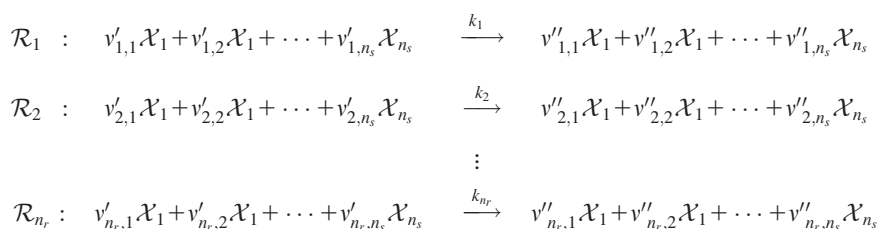
In this article, we first start with the formulation of the true (exact) likelihood of the data without any approximations. We demonstrate that for a very restricted class of models and data, the true likelihood may be computed and then numerically maximized to obtain maximum likelihood estimates (MLE). In further restricted cases, the posterior may be obtained by numerically integrating the relevant expressions. We then move on to implement and compare some of the Bayesian inference methods available in the literature. We consider only the methods that utilize the true stochastic chemical kinetic model (instead of using a diffusion approximation or a statistical substitute). These methods are based on Gibbs sampling and use either a Metropolis–Hastings (MH) step³⁰ or the uniformization^{32,33} technique. Finally, we describe a novel Bayesian inference method that we call conditional density importance sampling (CDIS), which uses importance sampling instead of MCMC techniques. We implement these three classes of methods (exact, MCMC and CDIS) on two examples from systems biology. The first example is a novel model of early viral gene expression, and the second example is a reduction of the RNA expression model studied by,^{20,22,23} which we call the gene on-off model. Each of these models exemplifies a quintessential feature of systems biology models that complicates parameter estimation. We use simulated datasets for each example and estimate the parameters using the different parameter estimation methods. Through these estimation exercises, we provide some general guidelines on how to utilize limited number of measurements to obtain maximum information. Finally, we compare all estimation methods and enumerate their strengths and weaknesses.

This article is organized as follows. We start with a brief review of stochastic chemical kinetic models and corresponding parameter estimation methods in the Background section. Following that, in the Examples section, we demonstrate the application of these parameter estimation methods on two systems biology models—early viral gene expression and gene on-off.

Background

Stochastic kinetic models

We provide only a brief description of stochastic chemical kinetics here. Consider a system of n_r reactions, denoted by $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_{n_r}$ and n_s species, denoted by $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_{n_s}$. The system of reactions is written in standard notation



in which, each reaction has an associated rate constant, k_i . We denote the set of rate constants using the $n_r \times 1$ vector, $\theta = [k_1 \ k_2 \ \dots \ k_{n_r}]^T$. Similar to the deterministic case, the $n_r \times n_s$ stoichiometric matrix, ν may be written as

$$\nu_{i,j} = \nu''_{i,j} - \nu'_{i,j}$$

Let $\mathbf{X}(t)$ be the $n_s \times 1$ state vector representing the number of molecules of each species at time t . Let \mathbf{R} be the $n_r \times 1$ vector denoting the number of times each reaction \mathcal{R}_i is fired during the time interval $[0, T]$. Given \mathbf{R} , the state of the system at time T may be obtained from the equation

$$\mathbf{X}(T) = \mathbf{X}(0) + \nu^T \mathbf{R} \quad (1)$$

But given $\mathbf{X}(T)$ and $\mathbf{X}(0)$, we can recover \mathbf{R} only if ν^T has full column rank. Equation 1 is not very different from the corresponding deterministic equation (see Rawlings and Ekerdt³⁴ p. 45, Eq. 2.30). Instead of using reaction rates and concentration of species, we use the number of reactions fired and number of molecules of each species. Note that both $\mathbf{X}(t)$ and \mathbf{R} are vector-valued random variables whose elements can assume only nonnegative integer values. At time t , each reaction \mathcal{R}_i has an associated reaction propensity defined as

$$h_i(\mathbf{X}(t), \theta) = k_i g_i(\mathbf{X}(t)) \quad (2)$$

For example, a bimolecular reaction of the form $A + B \xrightarrow{k} C$ would have reaction propensity $h(\mathbf{X}(t), \theta) = kA(t)B(t)$. Reaction propensity is similar to the rate of reaction in deterministic kinetics albeit with some differences as described in.² The total reaction propensity at time t is the sum of propensities of all reactions and is denoted by $h_0(\mathbf{X}(t), \theta) = \sum_{i=1}^{n_r} h_i(\mathbf{X}(t), \theta)$.

The set of all possible states of the system is called the state space, denoted by S . As the state is an integer-valued vector, the state space S is discrete containing $|S|$ states. Here, $|S|$ represents the number of states in the state space. Thus, the state vector forms a stochastic (or random) process $X = \{\mathbf{X}(t) : t \in [0, T]\}$ in a discrete state space and continuous time. This stochastic process is essentially a continuous-time discrete-space (homogeneous) Markov chain that has the following Markov property

$$\begin{aligned} P(\mathbf{X}(t+dt) = \mathbf{x} | \mathbf{X}(s) = \mathbf{x}(s), s \in [0, t]) \\ = P(\mathbf{X}(t+dt) = \mathbf{x} | \mathbf{X}(t) = \mathbf{x}(t)) \end{aligned} \quad (3)$$

Equation 3 can be converted to a system of ODEs by rearranging the terms and taking a limit of $dt \rightarrow 0$. One form of these ODEs is the Kolmogorov's forward equation

$$\frac{d}{dt} \mathbf{P}(t) = \mathbf{Q} \mathbf{P}(t) \quad (4)$$

$$\mathbf{P}(0) = \mathbf{I}_{|S| \times |S|}$$

$$(\mathbf{P}(t))_{i,j} = P(\mathbf{X}(s+t) = j | \mathbf{X}(s) = i), \quad i, j \in S \quad (5)$$

in which $\mathbf{P}(t)$ is an $|S| \times |S|$ transition matrix. Each element $(\mathbf{P}(t))_{i,j}$ represents the probability of the system being in state j at time $s+t$ given that the system was in state i at time t . The $|S| \times |S|$ matrix \mathbf{Q} is called the transition rate matrix. Elements of \mathbf{Q} are composed of the reaction propensities defined in Eq. 2. For a detailed description, we refer the reader to Ref. 30 (p. 137). Another form of the Kolmogorov forward equation, which is more common in the chemical and physical sciences literature, is the chemical

master equation (CME).^{4,5,11,22,35,36} The exact method and the uniformization technique, discussed in the next subsection, use Kolmogorov's forward equation as defined in Eq. 4.

The state of the system, $\mathbf{x}(t)$, remains constant over time until a reaction event occurs, thus forming a staircase plot of $\mathbf{x}(t)$ versus t . As a result, the stochastic process $X = \{\mathbf{X}(t) : t \in [0, T]\}$, which has uncountably infinite (time) dimensions, can be completely defined by recording only the reaction events. The simulation techniques^{2,6,7} exploit this characteristic by simulating reaction events or firings starting from an initial condition. For every reaction firing, two quantities are simulated from appropriate probability distributions—the time at which the reaction occurs and the index number of the reaction that occurs. Different trajectories of $\mathbf{X}(t)$ are generated using different random numbers. We denote one such trajectory or instance of X by the variable x . Note that x contains the complete reaction event data, which may be used to recreate the entire, continuously varying stochastic process. Unlike a simulation, laboratory experiments only measure the full (or partial) state of the system at discrete time points. These measurements constitute the discrete-time data and are collectively denoted by y . Note that y does not necessarily contain information about reaction events and is, therefore, a subset of the complete event data x .

Our aim is to estimate the parameter vector θ using a stochastic chemical kinetic model and discrete-time data y . In this article, we consider only the case of full measurements in which all species are measured at discrete time points. Let $\{s_0, s_1, \dots, s_m\}$ be the times at which measurements $\{y_0, y_1, \dots, y_m\}$ are available. Each measurement y_i is an $n_s \times 1$ vector. We also assume that the initial condition y_0 is known, that is, not random. Thus, the discrete-time data y can be defined as

$$y = \{y_0, y_1, \dots, y_m\} \quad (6)$$

Parameter estimation methods

We begin with a brief review of Bayesian inference as it applies to our problem. We assume that parameter vector θ is a random variable with a prior distribution $\pi(\theta)$. The posterior distribution of θ , denoted by $\pi(\theta|y)$, represents the updated distribution of θ after taking the data y into account. The prior and posterior are related via Bayes' rule

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \propto \pi(y|\theta)\pi(\theta) \quad (7)$$

in which $\pi(y|\theta)$ denotes the likelihood of data. Throughout this article, we use the following gamma prior

$$\pi(\theta) = \prod_{i=1}^{n_r} \pi(k_i) \quad (8)$$

$$\text{in which, } \pi(k_i) = \frac{b_i^{a_i}}{\Gamma(a_i)} k_i^{a_i-1} e^{-b_i k_i} \quad \text{or} \quad K_i \sim Ga(a_i, b_i) \quad (9)$$

We intend to obtain the posterior $\pi(\theta|y)$, or an approximation of it, using the gamma prior and the likelihood. MAP estimates, $\hat{\theta}$, may then be obtained by maximizing the posterior with respect to θ , either analytically or numerically. Unlike Bayesian estimation, MLE maximizes the likelihood $\pi(y|\theta)$ with respect to θ . We denote the estimates obtained by MLE as $\hat{\theta}_{MLE}$.

Exact Method. The exact method obtains the likelihood by solving Kolmogorov's forward equation described in (4) for a given value of θ , either numerically or analytically. The likelihood may then be maximized to obtain the MLE estimates $\hat{\theta}_{\text{MLE}}$, or the posterior may be obtained up to a constant, using Eq. 7. The likelihood $\pi(y|\theta)$ may be obtained as follows

$$\begin{aligned}\pi(y|\theta) &= P(\mathbf{X}(s_0)=y_0, \mathbf{X}(s_1)=y_1, \dots, \mathbf{X}(s_m)=y_m|\theta) \\ &= P(y_0, y_1, \dots, y_m|\theta) = P(y_m|y_{m-1}, y_{m-2}, \dots, y_0, \theta) \\ &= \prod_{i=1}^m P(y_i|y_{i-1}, \theta) P(y_0|\theta) \\ &= \prod_{i=1}^m P(y_i|y_{i-1}, \theta)\end{aligned}\quad (10)$$

Note that the term $P(y_i|y_{i-1}, \theta)$ is nothing but the (y_{i-1}, y_i) -element of the matrix $\mathbf{P}(s_{i+1}-s_i)$, which may be obtained by solving Eq. 4.

The solution to Eq. 4 requires a matrix exponentiation of $(s_{i+1}-s_i)\mathbf{Q}$, which becomes computationally expensive if the number of states, $|S|$, is large. As \mathbf{Q} depends on θ , at least one matrix exponentiation is required to calculate the likelihood value for every θ . As many as m matrix exponentiations may be required if the time points $\{s_0, s_1, \dots, s_m\}$ are not equally spaced. While some research has been done to solve the CME (and equivalently Kolmogorov's forward equation) analytically under special cases,³⁷ it is usually not possible to do so. This means that the likelihood value at a given θ has to be evaluated numerically which further increases the computational cost to obtain MLE estimates. The MAP estimates may be obtained by maximizing the product of likelihood and prior at a similar computational cost. To completely determine the posterior as a distribution, we require the normalizing constant $\pi(y)$, which is given by

$$\pi(y) = \int_0^\infty \pi(y|\theta) \pi(\theta) d\theta \quad (11)$$

In general, the integration in Eq. 11 is not analytically possible and numerical methods have to be used. Numerical integration of Eq. 11 is usually tractable when the number of parameters n_r is small (less than three). The gene on-off example in Examples section is one such case. In other cases, for example, early viral gene expression in the Examples section, the state space S is unbounded making \mathbf{Q} infinite dimensional and an analytical solution to Eq. 4 is essential to be able to use the exact method.

MCMC Methods. The MCMC methods and the CDIS method described in this section are in the same vein and utilize the same analytical expressions that are available under special circumstances. We begin with the general formulation of the estimation problem and describe relevant analytical expressions.

Given the complete event data x (instead of the discrete measurement data y), it is possible to write down the likelihood analytically [30, p. 279]

$$\pi(x|\theta) = \prod_{j=1}^n h_{n_j}(\mathbf{x}(t_{j-1}), \theta) \exp \left\{ - \int_{s_0}^{s_m} h_0(\mathbf{x}(t), \theta) dt \right\} \quad (12)$$

in which

- n represents the total number of reaction events occurred in the time interval $[s_0, s_m]$,
- n_j represents the index number of the j th reaction event,
- t_j represents the time at which j th reaction event occurred.

Note that Eq. 12 computes the likelihood of the continuously varying stochastic process using only the information about reaction events. When the reaction propensity is separable, that is, as described in Eq. 2, the likelihood expression becomes

$$\pi(x|\theta) = g_{\text{prod}} \prod_{i=1}^{n_r} k_i^{r_i} e^{-G_i k_i} \propto \prod_{i=1}^{n_r} k_i^{r_i} e^{-G_i k_i} \quad (13)$$

in which

$$g_{\text{prod}} = \prod_{j=1}^n g_{n_j}(\mathbf{x}(t_{j-1})) \quad G_i = \int_{s_0}^{s_m} g_i(\mathbf{x}(t)) dt \quad (14)$$

and r_i is the i th element of \mathbf{r} representing the number of times reaction \mathcal{R}_i occurs in the time interval $[s_0, s_m]$. This expression of likelihood motivates the use of the gamma prior defined in Eq. 8. The posterior given complete reaction event data may be written analytically as

$$\pi(\theta|x) = \prod_{i=1}^{n_r} \pi(k_i|x) \quad (15)$$

in which

$$\begin{aligned}\pi(k_i|x) &= \frac{(b_i + G_i)^{a_i + r_i}}{\Gamma(a_i + r_i)} k_i^{a_i + r_i - 1} e^{-(b_i + G_i)k_i} \quad \text{or} \\ K_i|x &\sim \text{Ga}(a_i + r_i, b_i + G_i)\end{aligned}\quad (16)$$

Thus, the prior and complete data posterior are conjugate, that is, they come from the same family of distributions.

The problem is that we have the discrete-time (measurement) data y instead of the complete (reaction) event data x . As $y \subseteq x$, an obvious approach is to first appropriately simulate and average over the hidden or missing data represented by $x \setminus y$, and then infer the parameters θ using the "averaged" x . Both the MCMC and CDIS methods essentially perform this averaging in very different ways. We will now describe the MCMC methods and delay the description of CDIS until later in this subsection.

MCMC methods aim to obtain the joint distribution of θ and x given y , that is, $\pi(\theta, x|y)$ by using Gibbs sampling [30, p. 255]. A basic Gibbs-style algorithm may be written as:

1. Sample $x \sim \pi(x|y)$
2. Sample $\theta \sim \pi(\theta|x, y) = \pi(\theta|x)$ using Eq. 16
3. Sample $x \sim \pi(x|\theta, y)$
4. Return to Step 2

This algorithm may be iterated to produce samples of θ that can then be binned to produce a histogram representing the posterior $\pi(\theta|y)$. Note that the samples of θ and x generated by this algorithm come from the true joint distribution $\pi(\theta, x|y)$. There are two problems with this method. First, the distribution $\pi(x|y)$ is not available. So, we initialize the iterative procedure by replacing Step 1 with any method (not necessarily statistical) that can generate a sample of x that satisfies y . Obviously, when the initial x does not come from the true distribution $\pi(x|y)$, the samples of θ and x are no longer distributed as $\pi(\theta, x|y)$. In this case, the samples of

(θ, x) form a Markov chain whose stationary distribution is $\pi(\theta, x|y)$. Thus, the Markov chain can be run until stationarity is achieved and then samples of θ may be generated to produce a histogram. The second problem is to sample x from $\pi(x|\theta, y)$. This problem can either be solved using endpoint-conditioned simulation methods³³ or using a MH step [30, p. 282].

Endpoint-conditioned simulation methods, such as uniformization,³³ can generate samples of x from the true conditional distribution $\pi(x|\theta, y)$. This method, which we refer to as Gibbs-uniformization, was used by³² to estimate parameters in epidemic models. Similar to the exact method, the uniformization technique uses the solution to Kolmogorov's forward equation and, therefore, suffers from similar associated issues of a large (or unbounded) state space. Other endpoint-conditioned simulation methods such as direct sampling³³ or bisection sampling³⁸ may be used to produce similar estimation methods.

Using a MH step, [30, p. 282] makes the overall algorithm a Metropolis-within-Gibbs-style algorithm. The MH step generates a proposed sample of x from an approximate conditional distribution $\pi_A(x|\theta, y)$ instead of the true conditional, and then, accepts or rejects the proposed sample x based on a suitably defined acceptance probability. This class of estimation methods was proposed by Wilkinson and coworkers.^{26,28–30} There are two major variations of this overall scheme—RJMCMC method³¹ and the block updating method.^{30,31} In this article, we consider only the block updating method as described in Ref. 30 [pp. 281–287], which we call MCMC-MH.

CDIS Method. We present here an alternative and novel method, CDIS, to perform the “averaging” over x as discussed previously in this section. Instead of attempting to obtain the joint posterior distribution of (θ, x) given y , we aim to obtain only the posterior $\pi(\theta|y)$ using importance sampling

$$\begin{aligned}\pi(\theta|y) &= \int_x \pi(\theta, x|y) dx = \int_x \pi(\theta|x) \pi(x|y) dx \\ &\approx \sum_{k=1}^N \pi(\theta|x_k) \frac{\pi(x_k|y)}{q(x_k)} \\ &= \frac{1}{\pi(y)} \sum_{k=1}^N \pi(\theta|x_k) \frac{\pi(y|x_k) \pi(x_k)}{q(x_k)}\end{aligned}\quad (17)$$

in which, $q(x)$ represents the importance or proposal distribution. Instead of generating samples of x from the target distribution $\pi(x|y)$, we generate samples from the importance distribution. Note that as $y \subseteq x$, the reverse conditional $\pi(y|x)$ is either 1 (when x agrees with y) or zero (otherwise). The distribution of complete event data $\pi(x)$ may be obtained using Bayes' rule as follows

$$\pi(x) = \frac{\pi(x|\theta) \pi(\theta)}{\pi(\theta|x)} = g_{\text{prod}} \prod_{i=1}^{n_r} \frac{\Gamma(a_i + r_i)}{\Gamma(a_i)} \frac{b_i^{a_i}}{(b_i + G_i)^{a_i + r_i}} \quad (18)$$

Using Eqs. 17, 13, and 18, the posterior may be approximated by a weighted sum of gamma distributions

$$\pi(\theta|y) \approx \hat{\pi}(\theta|y) = \sum_{k=1}^N w_k \pi(\theta|x_k) \quad (19)$$

in which the weights w_k are given by

$$w_k = \frac{\frac{\pi(y|x_k) \pi(x_k)}{q(x_k)}}{\sum_{k=1}^N \frac{\pi(y|x_k) \pi(x_k)}{q(x_k)}} \quad (20)$$

Note that the complete data likelihood $\pi(\theta|x)$ is known analytically in Eq. 15, and the weights w_k can be numerically computed, thus, providing a semianalytical expression for the joint posterior in Eq. 19.

As is usually the case with importance sampling, the choice of the importance function is critical to numerical efficiency. The importance function $q(x)$ should “mimic” the target distribution $\pi(x|y)$. In this article, we present the importance sampling procedure in brief and refer the reader to Ref. 39 for a detailed description with statistical analysis. To obtain $q(x)$, we start with the target distribution

$$\begin{aligned}\pi(x|y) &= \int_{\theta} \pi(x, \theta|y) d\theta = \int_{\theta} \pi(x|\theta, y) \pi(\theta|y) d\theta \\ &\approx \int_{\theta} \pi(x|\theta, y) \pi(\theta) d\theta\end{aligned}\quad (21)$$

As we discussed previously in this section, x contains the information about the time and the index number of reaction events that occur in the time interval $[s_0, s_m]$. To simulate x , we first generate the random variable \mathbf{R} , using the distribution $q(\mathbf{r})$. Conditioned on \mathbf{r} , we generate reaction sequence $\{n_1, n_2, \dots, n_n\}$ in which n_j represents the index number of j th reaction event and $n = \sum_{i=1}^{n_r} r_i$. Finally, conditioned on \mathbf{r} and $\{n_1, n_2, \dots, n_n\}$, we generate a set of reaction times $\{t_1, t_2, \dots, t_n\}$, which represent the times corresponding to the reaction events. Thus, the $q(x)$ is given by

$$q(x) = q(\mathbf{r}) q(n_1, n_2, \dots, n_n | \mathbf{r}) q(t_1, t_2, \dots, t_n | n_1, n_2, \dots, n_n, \mathbf{r}) \quad (22)$$

Note that both Gibbs-uniformization and MCMC-MH methods generate samples of x for a given θ , from the (target) distribution $\pi(x|\theta, y)$. This requires an iteration over θ and produces the posterior $\pi(\theta|y)$ as a histogram. By contrast, CDIS method generates samples of x from the (target) distribution $\pi(x|y)$ in which θ has been integrated out. Thus, the CDIS method produces a semianalytical posterior distribution that can be evaluated for every θ value.

Examples

The reaction kinetics in systems biology (whether deterministic or stochastic) differ from the reaction kinetics in chemistry in one large respect—systems biology models usually violate mass balance.^{17,20–26,29–32} Such reaction kinetics can result in an unbounded number of molecules (usually as a result of a synthesis reaction) causing the state space S to become unbounded as well. Violation of mass balance is an ease-of-use approximation in which we do not explicitly model species (e.g., amino acids and nucleic acids) that (1) are present in large amounts, (2) are not expected to not vary appreciably, and (3) are not critical to the reaction system we intend to model. Balancing the reactions by explicitly modeling “missing” species does not improve the situation. The state space, S , remains prohibitively large (although bounded) due to the species present in large amounts, and the increased number of species causes an unnecessary increase in model complexity and computation.

Table 1. True Rate Constants and Gamma Prior Parameters

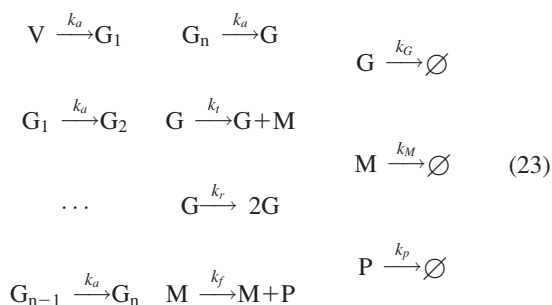
Reactions	Parameters	True Values		Prior Parameters	
		θ_0	$\hat{\theta}_{\text{prior}}$	a	b
Activation	k_a	0.15	10	1.01	0.0010
Transcription	k_t	0.02	20	1.01	0.0005
Replication	k_r	0.05	1	1.01	0.0100
Translation	k_f	1.00	30	1.01	0.0003

Unboundedness of state space, S , usually rules out the application of exact and Gibbs-uniformization methods, except when an analytical solution to Eq. 4 is available. Our first example, early viral gene expression, in the Examples section has this feature.

Another common feature that complicates parameter estimation is cyclical kinetics. Cyclical kinetics includes reversible reactions as well as oscillatory kinetics such as the Lotka–Volterra model. We define cyclical kinetics as the set of reactions whose transposed stoichiometric matrix v^T has less than full column rank. As discussed in the Background section, when v^T does not have full column rank, Eq. 1 cannot be solved uniquely for \mathbf{R} , thus, resulting in at least one degree of freedom. In such a case, a large, possibly infinite, number of (hidden) reaction events may occur between two measurements. As a result, the amount of hidden data to be “averaged” over is much larger, which demands more computational effort. The gene on-off example in the Examples section has this feature.

Early viral gene expression model

The motivation for this example comes from the virus infection experiments performed by Timm and Yin.⁴⁰ The following model describes the viral infection of a single baby hamster kidney (BHK-21) cell by VSV. The rationale behind this reaction mechanism comes from both the VSV-BHK biology and the characteristics of the data

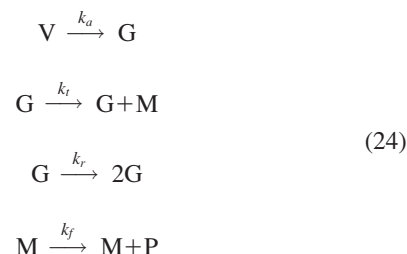


The viral infection process takes place in five distinct stages⁴¹: adsorption to the host cell and entry, uncoating of the genome, transcription and translation, genome replication, assembly and release of virus progeny. The first two stages, namely, adsorption/entry and uncoating must occur before the viral genome undergoes transcription or replication. A delay is observed before viral protein begins to appear.

In (23), V represents unactivated viral genome and G represents activated viral genome. The species $\text{G}_1, \text{G}_2, \dots, \text{G}_n$ represent various intermediate (or delay) states of the viral genome. Unactivated viral genome, V, has to undergo adsorption, entry, and uncoating before it is ready for transcription and replication. The reactions in (23) converting V to G represent these initial stages of adsorption, entry, and uncoating. The species M and P

represent Red Fluorescent Protein (RFP) mRNA and red fluorescent protein, respectively. The transcription reaction creates more mRNA, M from activated genome G. Translation creates more red fluorescent protein, P from mRNA, M. Observed RFP signal is proportional to the amount of red fluorescent protein P present in the cell. Here, any delays associated with maturation of the RFP signal are embedded in the translation rate.

During the early stages of infection, the degradation of G, M, and P is not important. Therefore, we further reduce the model in (23) to represent only the early stages of infection and remove the delay states. This leads to the following reduced model

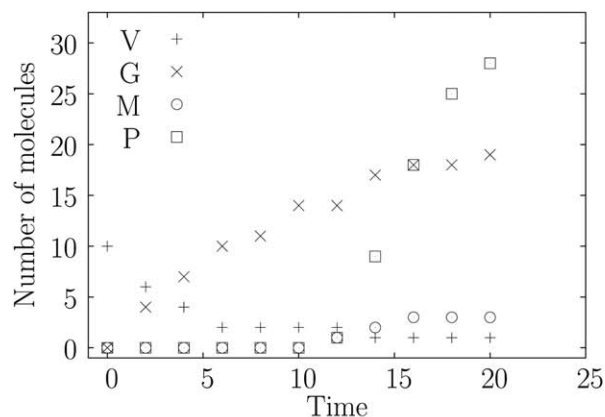


Using true parameter values θ_0 as shown in Table 1 and an initial condition of $\mathbf{X}(0) = [\text{V}_0 \text{ G}_0 \text{ M}_0 \text{ P}_0]^T = [10 \ 0 \ 0 \ 0]^T$ (corresponding to a multiplicity of infection or MOI of 10), we generate six simulated trajectories with Gillespie’s direct method.² These trajectories, shown in Figure 1, use the same random numbers but have different sampling frequency (Δt) and different number of time points (m). Note that Figures 1a, b span the same time interval of [0, 20] but Figure 1b has 20 times larger measurement frequency. The same is true for the next two rows of Figure 1.

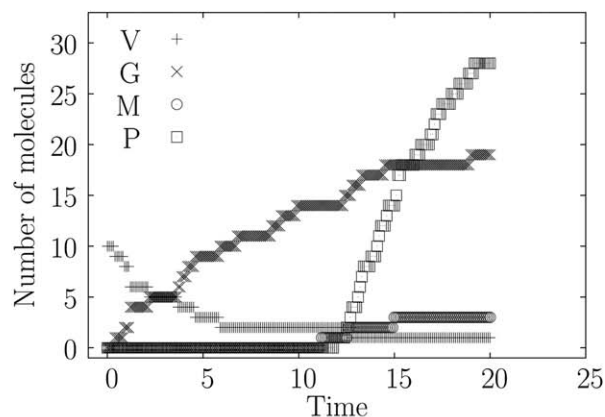
As we can see in (24), all reactions except the activation reaction violate mass balance. This rules out the application of exact and Gibbs-uniformization methods and we only compare the MCMC-MH and CDIS methods. We use a gamma prior as described in Eq. 8 with shape (a) and rate (b) parameters shown in Table 1. The mode (or peak) of the gamma prior, denoted by $\hat{\theta}_{\text{prior}}$ is also shown in Table 1. Note that $\hat{\theta}_{\text{prior}}$ is chosen to be very different than θ_0 so as to not bias the posterior toward the true values. The shape parameter is chosen so that the prior is essentially uniform over a large range of parameter values.

We estimate the parameters for each of the six trajectories using both CDIS (with $N = 1000$ samples) and MCMC-MH methods (with $N = 10,000$ steps). The resulting parameter estimates (MAP estimates) are shown in Table 2. The marginal posteriors $\pi(k_i|y)$ are shown in Figure 2. As discussed before, CDIS produces a semianalytical expression for the marginal posterior while the MCMC-MH method produces samples of θ , which are then binned to produce the histogram. Parameter estimates for the CDIS method are obtained by numerical optimization. The estimates for MCMC-MH method are obtained by simply finding the bin with the largest probability (or height). Histogram estimation suffers from the problem of bias-variance tradeoff [42, p. 303]. As a result, MCMC-MH estimates have an inherent error corresponding to the width of the bin ($\sim 10^{-2} - 10^{-5}$ for this example). CDIS estimates in contrast are accurate up to any desired level of accuracy ($\sim 10^{-10}$ due to machine precision).

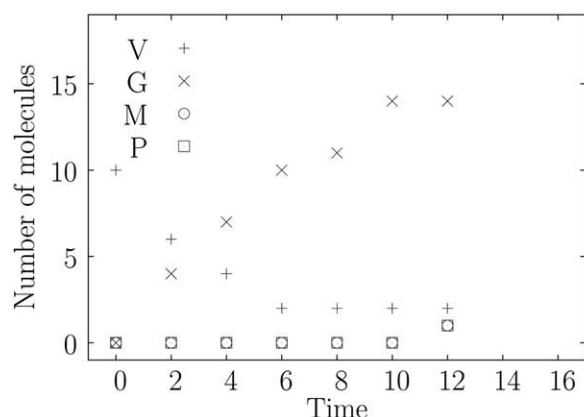
We begin by comparing the results for trajectories in Figures 1a, b. Figures 2a, b show the corresponding marginal posteriors and priors. In comparison to the posteriors, the priors are essentially uniform, indicating that the priors have little effect on the MAP estimates. The CDIS and



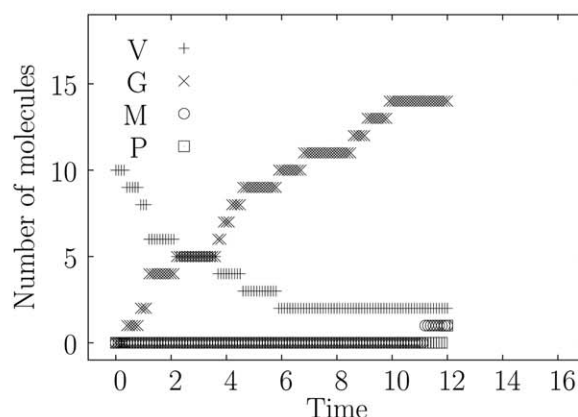
(a) $m = 10, \Delta s = 2$



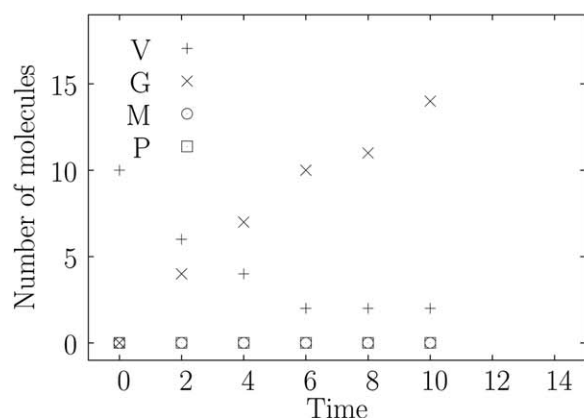
(b) $m = 200, \Delta s = 0.1$



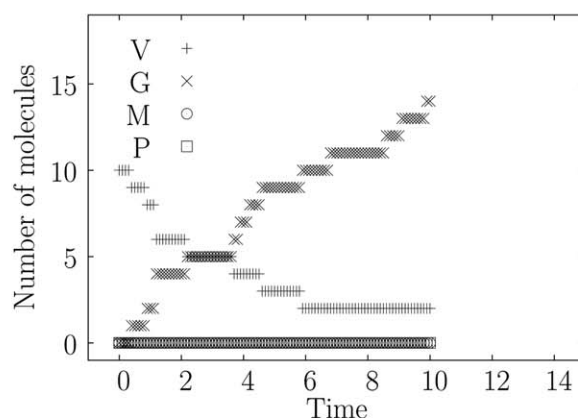
(c) $m = 6, \Delta s = 2$



(d) $m = 120, \Delta s = 0.1$



(e) $m = 5, \Delta s = 2$



(f) $m = 100, \Delta s = 0.1$

Figure 1. (A-F) VSV early gene expression.

True parameter values of $\theta_0 = [k_a k_t k_r k_f]^T = [0.15 \ 0.02 \ 0.05 \ 1]^T$. Initial condition of $X(0) = [V_0 \ G_0 \ M_0 \ P_0]^T = [10 \ 0 \ 0 \ 0]^T$ corresponding to an MOI of 10. m = number of points, Δs = sampling time.

MCMC-MH methods agree quite well. The first two rows of Table 2 provide the parameter estimates. The parameter estimates $\hat{\theta}$ are close to the true values θ_0 indicating that the estimation methods are reasonably accurate. For any statistical estimation method, even the exact method, there is some estimator error (or bias) due to the finite amount of data. Thus, we do not expect the estimates in Table 2 to track the

true values. However, we expect the parameter estimates, $\hat{\theta}$ to converge as the number of samples (N) increases.

Note that increasing the number of measurements by a factor of 20 (from 1a to 1b) does appreciably change the parameter estimates. Looking at the next trajectory in Figure 1c we see that the CDIS and MCMC-MH posteriors for translation rate constant, k_f , are different (Figure 1c). In this

Table 2. Parameter Estimates from CDIS and MCMC-MH

Dataset	CDIS ($N=1000$)	MCMC-MH ($N=10,000$)
(1a)	(0.1674, 0.0124, 0.0411, 1.2767)	(0.1457, 0.0117, 0.0380, 1.3568)
(1b)	(0.1716, 0.0122, 0.0407, 1.3163)	(0.1768, 0.0112, 0.0387, 1.3871)
(1c)	(0.1825, 0.0096, 0.0572, 0.8019)	(0.1907, 0.0082, 0.0597, 6.4541)
(1d)	(0.1832, 0.0095, 0.0567, 1.2069)	(0.1752, 0.0121, 0.0549, 1.1623)
(1e)	(0.1994, 0.0001, 0.0778, 30.0000)	(0.2012, 0.0006, 0.0700, 155.6734)
(1f)	(0.2018, 0.0001, 0.0770, 30.0000)	(0.2010, 0.0008, 0.0759, 166.1711)

True values $(k_{a0}, k_{r0}, k_{f0}) = (0.15, 0.02, 0.05, 1)$.

case, increasing the number of measurements causes the two posteriors to agree with each other (Figure 2d). Looking at the corresponding parameter estimates in Table 2, we can see that increasing the measurement frequency changes the estimates for k_f significantly but the other parameter estimates do not change. Finally, for the last two trajectories in Figures 1e, f, increased measurement frequency does not change the parameter estimates as well. However, for these trajectories, there are two interesting outcomes—(1) the parameter estimates for k_i and k_f are very different from their corresponding true values and (2) the posteriors for k_f are the same as the prior.

MCMC-MH and CDIS methods use the complete data posterior $\pi(\theta|x)$ (in Eqs. 15–16) to generate an approximation of the required posterior $\pi(\theta|y)$. Equation 16 offers some explanation about behavior of estimates. The marginal complete data posterior, $\pi(k_i|x)$ uses the information in x only through r_i and G_i . Here, r_i is the number of times reaction \mathcal{R}_i occurs in time interval $[s_0, s_m]$ and can be obtained uniquely by solving Eq. 1 because v^T has full column rank. Changing the measurement frequency while keeping the same end points y_0 and y_m does not change r_i . But increasing the measurement frequency does provide a better (lower variance) estimate of G_i (defined in Eq. 14). This observation explains why the increasing measurement frequency has a weak effect on the posterior and estimates.

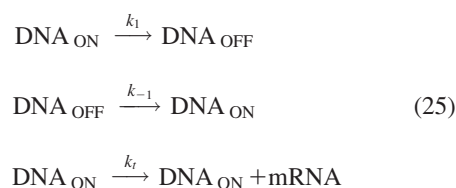
We now compare the data in Figures 1e, c, a. The trajectory in Figure 1e shows that the mRNA (M) and protein (P) measurements remain at zero and neither transcription nor translation reaction events occur during the given $[0, 5]$ time interval. As a result, every (event data) sample x generated by the CDIS or MCMC-MH method has $r_4=0$ and $G_4=0$. Looking back at Eq. 16, we can see that this dataset provides no information at all for the translation reaction. Consequently, the posterior of k_f is exactly equal to the posterior. While the CDIS method provides this information analytically, the MCMC-MH method provides this information through samples of k_f . In such a case, the model in (24) should be further reduced by removing the translation reaction. While r_2 is also equal to zero, $G_2 > 0$, which means that there is some information available regarding the transcription rate constant k_t . Consequently, the estimate \hat{k}_t in Table 2, while being very different than $k_{t,0}$, is not equal to $\hat{k}_{t,\text{prior}}$. Figures 1c, a have larger (nonzero; r_i, G_i) values and provide better estimates.

In Table 3, we present the computational time spent in terms of seconds taken to generate one sample of x for both methods. All simulations were performed using Octave on an Intel(R) Core(TM) 2 Quad 2.66GHz, 8GB RAM machine running Ubuntu 13.04. For both methods, the computational time per sample is directly proportional to the number of time points, m and is (roughly) inversely proportional to the sampling interval, Δs . Table 3 shows that when m is low (in

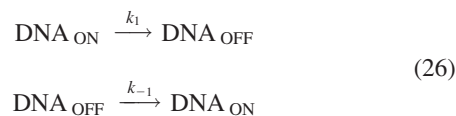
1a, c, e), the CDIS method takes much longer to generate one sample of x than MCMC-MH. However, when m is high (in 1b, d, f), MCMC-MH is slower than CDIS. Irrespective of the per sample computational time, CDIS requires at least 10 times fewer samples than MCMC-MH. In fact, CDIS method produces a semianalytical posterior with just one sample while a large (~ 1000) samples are required to obtain a histogram using MCMC methods. Even further, a significant number of samples (~ 10 –50%) have to be burned-in while using MCMC methods to ensure convergence of the Markov chain to stationarity. The CDIS method, by contrast, may have problems with convergence if the importance distribution $q(x)$ is not “close enough” to the target distribution $\pi(x|y)$. We have not encountered convergence issues for this example (see Supporting Information).

Gene on-off

The RNA expression model was first proposed by Golding et al.²⁰ and later studied by Reinker et al.²³ and Poovathingal and Gunawan.²² This model, as shown in reactions (25), has three species and three reactions

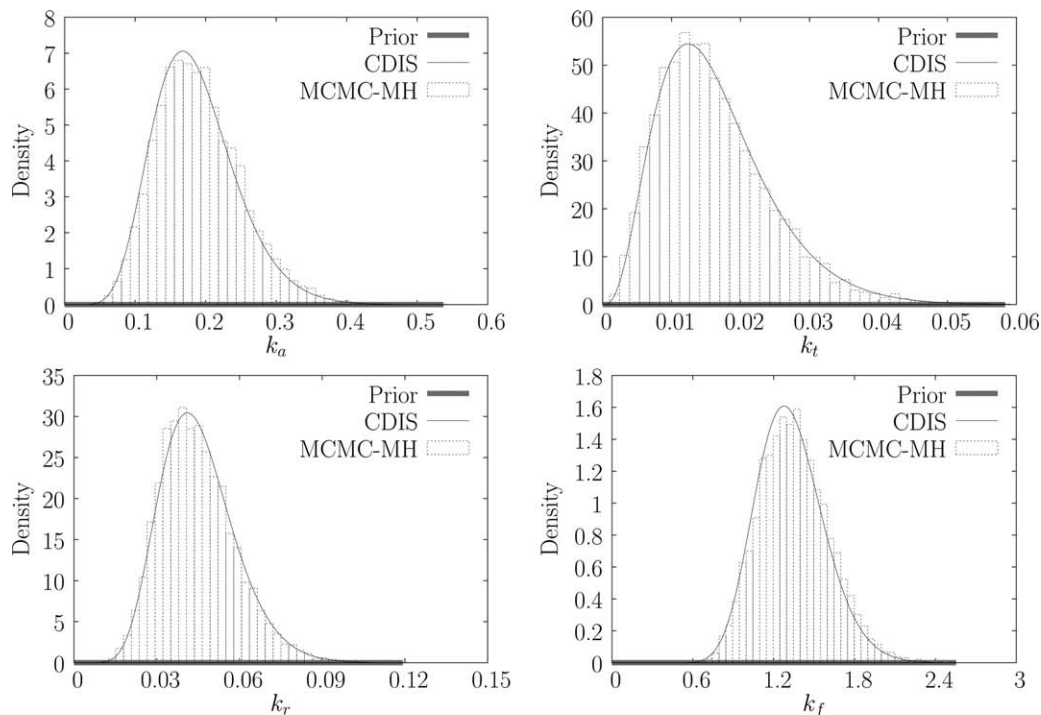


As discovered by Golding et al.,²⁰ the number of RNA molecules in a single cell was around 0–10. This simple model also violates mass balance, which makes the exact and Gibbs-uniformization methods inapplicable. To demonstrate the exact and Gibbs-uniformization methods and compare them to MCMC-MH and CDIS methods, we consider only the first two reactions in the model. This simplification creates the reduced, gene on-off model presented in reactions (26)

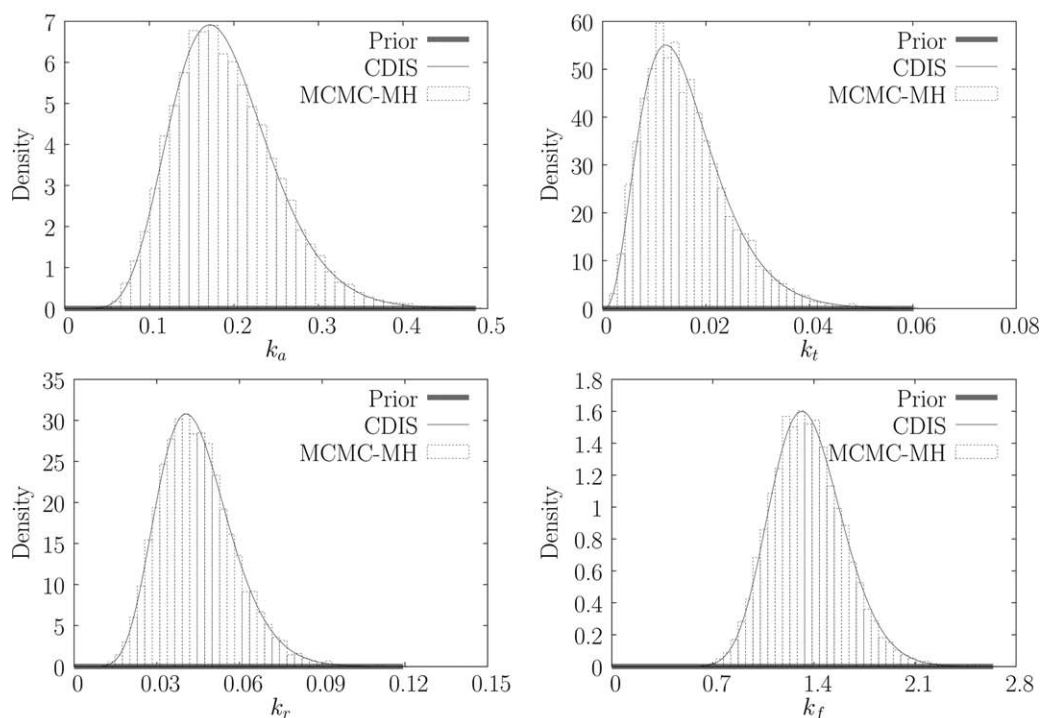


While this model is no longer suited to study mRNA expression, it can be used to study the switching—activation (ON) and deactivation (OFF)—of a gene. As this model now has only two parameters to be estimated, we can plot likelihood and joint posterior values, which is another motivation for choosing this reduced model.

We generate six simulated trajectories using true parameter values $\theta_0 = [k_1 \ k_{-1}]^T = [0.03 \ 0.01]^T$ with Gillespie's direct method,² as shown in Figure 3. Trajectories in the first



(a) $m = 10, \Delta s = 2$

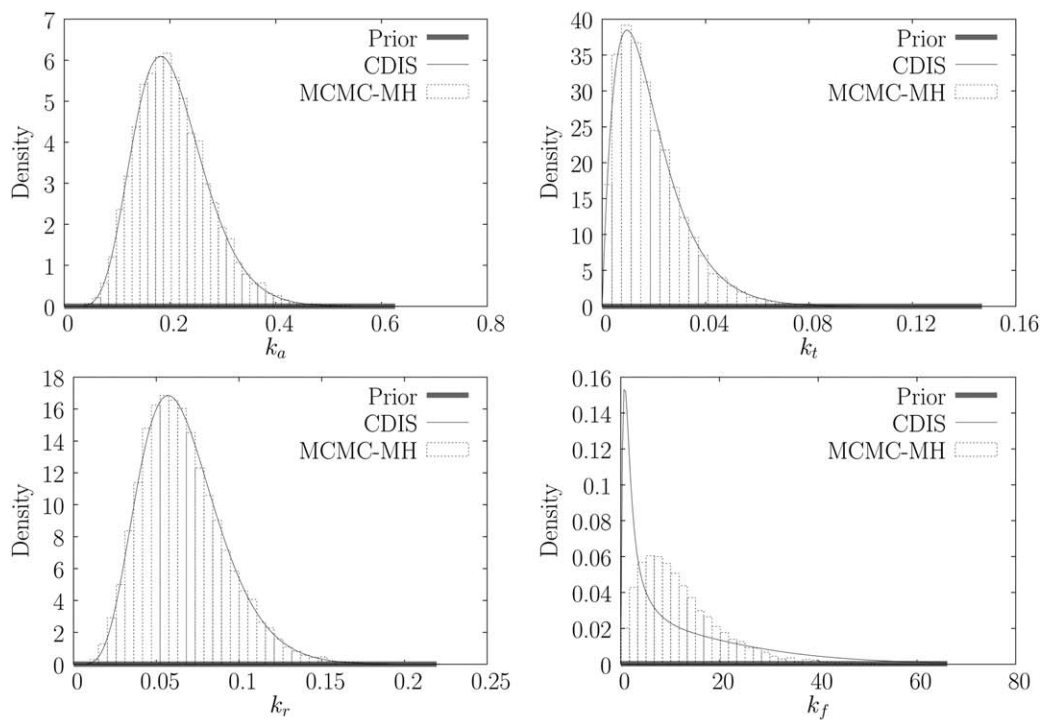


(b) $m = 200, \Delta s = 0.1$

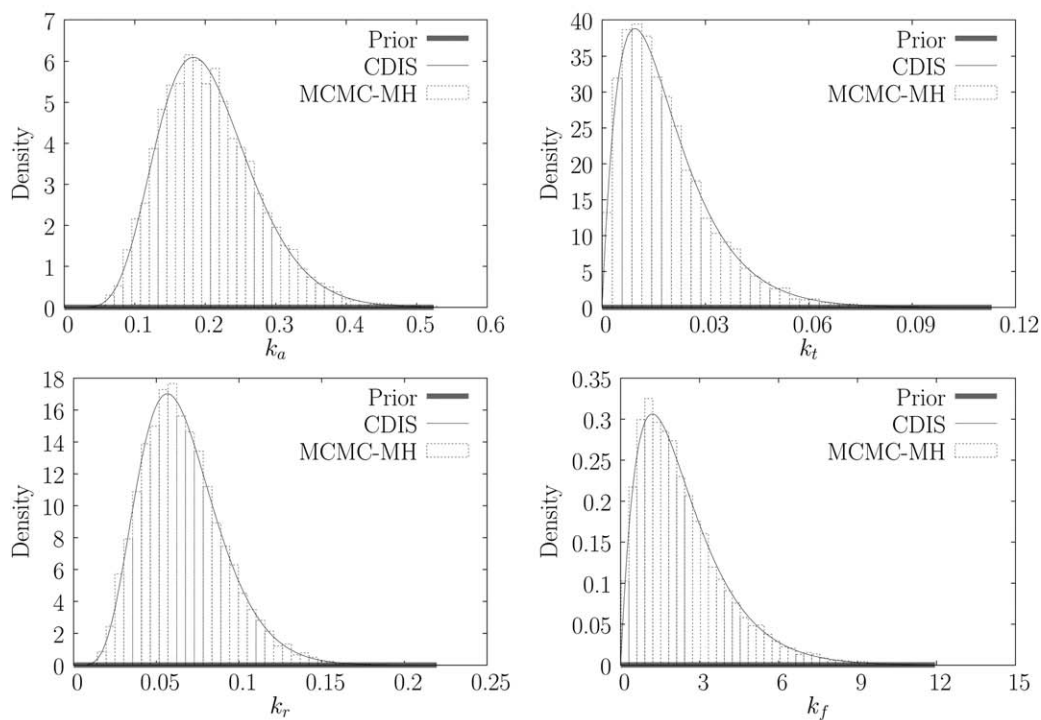
Figure 2. Marginal priors and posteriors obtained using CDIS and MCMC-MH.

column (Figures 3a, c, e) are simulated using an initial condition of $\mathbf{X}(0) = [\text{DNA}_{\text{ON},0} \quad \text{DNA}_{\text{OFF},0}]^T = [1 \quad 0]^T$, which corresponds to only one copy of the gene. The second column (Figures 3b, d, f) is simulated using $\mathbf{X}(0) = [\text{DNA}_{\text{ON},0} \quad \text{DNA}_{\text{OFF},0}]^T = [5 \quad 4]^T$, which corresponds to nine copies of the gene.

We begin by computing the likelihood $\pi(y|\theta)$ for each trajectory using the exact method as described in the Background section. The resulting three-dimensional (3-D) likelihood plots, corresponding to each of the trajectory, are presented in Figure 4. We can see that for the data in Figure 3a, the likelihood is essentially flat when k_{-1} is small and k_1



(c) $m = 10, \Delta s = 2$



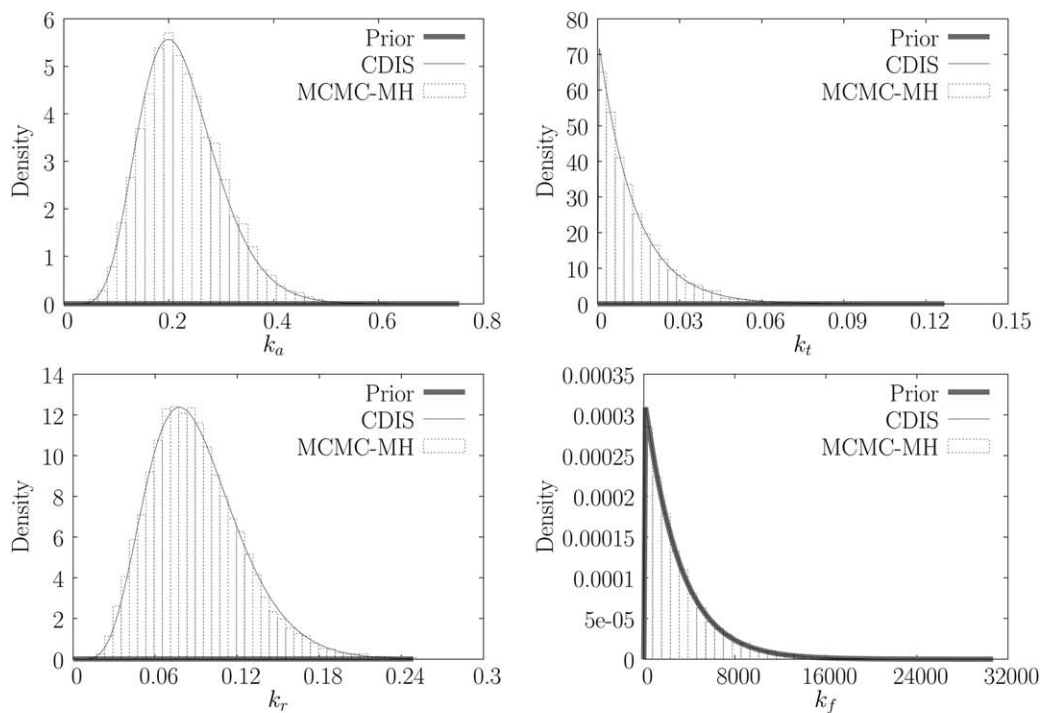
(d) $m = 200, \Delta s = 0.1$

Figure 2. (Continued)

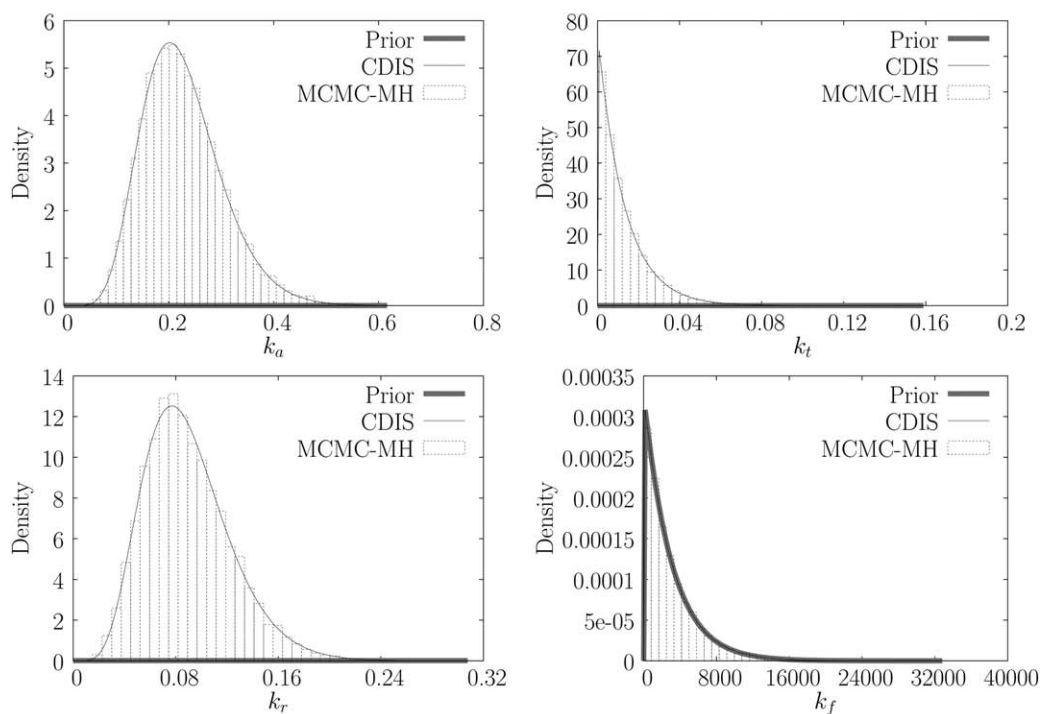
is large enough. The likelihood plots for Figure 3b, c are flat when k_{-1} is small but have a peak in k_1 . The data in Figures 3d, e produce likelihood curves, which are peaked in both k_1 and k_{-1} while the data in Figure 3f the most sharply peaked likelihood surface.

Intuitively, the likelihood in Figure 4a seems reasonable. The corresponding trajectory in Figure 3a may be recreated,

with probability 1, if we set $k_1 = \infty$ and $k_{-1} = 0$, that is, $\pi(y|\theta = [\infty \ 0]^T) = 1$. Therefore, the only information we can obtain is that $k_1 \gg k_{-1}$; we cannot obtain independent estimates of both parameters. Figure 3b has the same number of measurements as Figure 3a but $k_1 = \infty$ is no longer possible because we observe that \mathcal{R}_1 does not occur until $t = 2$. This observation provides some information about k_1 . As \mathcal{R}_2



(e) $m = 10$, $\Delta s = 2$



(f) $m = 200$, $\Delta s = 0.1$

Figure 2. (Continued)

does not occur in Figure 3b, it is still possible for k_{-1} to be zero. Data in Figure 3c has a large number of finely sampled measurements ($m=800$, $\Delta s=0.1$) but it still does not provide more information than Figure 3b, which has far fewer measurements. Figure 3c indicates that the gene switched off (i.e., \mathcal{R}_1 occurred) once around $t=15$. While the increased measurement frequency provides an accurate time of the switch

(compared to Figures 3a, b), the trajectory can still be simulated when $k_{-1}=0$. Thus, Figure 3c provides no information about \mathcal{R}_2 . In fact, a reduced model consisting of \mathcal{R}_1 alone may be used to explain this data. Figure 3d has 10 times fewer measurements than Figure 3c and spans the same time interval of $[0, 80]$, but it provides a better estimate of k_{-1} . This is because both \mathcal{R}_1 and \mathcal{R}_2 occur often. Figure 3e also

Table 3. Sampling Time (Seconds per Sample) for CDIS and MCMC-MH

Dataset	CDIS ($N=1000$)	MCMC-MH ($N=10,000$)
(1a)	0.54	0.12
(1b)	0.54	1.35
(1c)	0.30	0.06
(1d)	0.37	0.83
(1e)	0.30	0.05
(1f)	0.35	0.66

provides better estimates for the same reason. Note that the data in Figures 3d, e do not allow the possibility of $k_{-1}=0$. Figure 1f provides the sharpest likelihood surface (and the best estimates) among all trajectories in spite of having a sampling time of $\Delta s=20$.

Unlike the previous example, v^T for this example does not have full column rank. This means, as we intuitively expect, that we cannot uniquely obtain the number of times each reaction occurred (i.e., r_i) between two measurements. For example, in Figure 3a, the number of times \mathcal{R}_1 and \mathcal{R}_2 occurred between $t=0$ and $t=20$ is given by $(r_1, r_2)=(n+1, n)$, $n=0, 1, 2, \dots$. Therefore, we define $r_{i,\min}$ as the observed (or minimum) number of times a reaction \mathcal{R}_i must occur. As we can see, for Figures 3a, b, c, $(r_{1,\min}, r_{2,\min})=(1, 0)$, which indicates that these datasets do not provide information about k_{-1} . Figures 3d, e, f have larger (nonzero) values of $(r_{1,\min}, r_{2,\min})$ and provide better estimates.

We compare other estimation methods with the exact method for the data in Figure 3d. The joint posterior for all four estimation methods is presented in Figure 5. Parameter estimates obtained by (exact) maximum likelihood estimation and other methods are provided in Table 4. The estimates obtained by exact Bayesian estimation are close to the exact MLE estimates, which indicate that the prior does not bias the estimation. The sampling times are also presented in Table 4. Gibbs-uniformization method performs the worst due to slow mixing of the Markov chain (see Supporting Information). Furthermore, the uniformization step is extremely slow, which limits the number of Markov chain steps that can be used. Detailed analysis of computation time required by uniformization and other similar methods may be found in Ref. 33. As we can see from Figure 5d, 1000 Markov chain steps were not enough to produce a 3-D histogram. Consequently, the Gibbs-uniformization estimates are not accurate. Gibbs-uniformization must be run for a much larger number of steps to obtain reliable estimates. By contrast, CDIS ($N=100$ samples) and MCMC-MH ($N=1000$ steps) estimates are close to the exact (Bayesian) estimates. A burn-in of 100 steps was used for MCMC-MH method. Variance in CDIS estimates (using $N=100$ samples) was lower than the variance in MCMC-MH estimates (using $N=1000$ steps). Reducing the number of Markov chain steps used by the MCMC-MH method increases the variance in the resulting histogram and parameter estimates. Thus, the CDIS method requires 10 times fewer numbers of samples compared to the MCMC-MH method to obtain the same accuracy in estimates. The computational expense in terms of seconds/sample is higher for CDIS compared to the MCMC-MH method but the overall computation time is still $\sim 85\%$ lower for CDIS method than MCMC-MH method. In this article, we implemented the block update version of

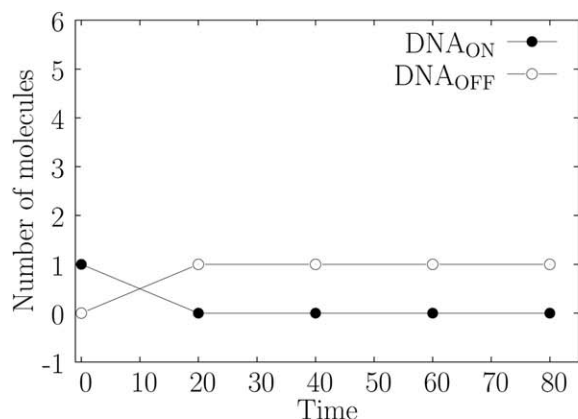
the Metropolis-within-Gibbs-style estimation methods (see MCMC methods in the Background section). It is possible that the RJMCMC method³¹ may perform better. For cyclical kinetics, both the CDIS and the Metropolis-within-Gibbs-style methods require a tuning parameter to generate samples of r_i . We followed the recommended tuning parameter as suggested by Boys et al.³¹ and Wilkinson.³⁰ More details are provided in the Supporting Information. The marginal posteriors for both parameters are presented in Figure 6. We used $a=[1.01 \ 1.01]^T$, $b=[0.00067 \ 0.00125]^T$ as the parameters for gamma prior.

Experimental and Model Design

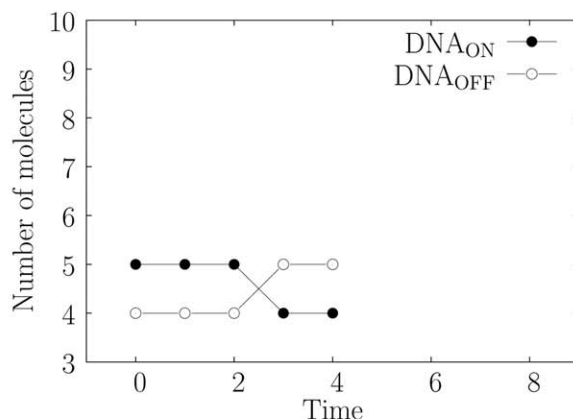
Only a limited number of measurements can be performed in the laboratory due to resource limitations. The analysis presented in the previous section may be used to develop general guidelines on both experimental and model design that will help in parameter estimation. As we have shown, performing more measurements at a higher frequency does not necessarily improve estimates of some parameters. In both examples, we obtained inaccurate estimates of a rate constant k_i when the corresponding reaction \mathcal{R}_i was not observed to occur, that is, $r_{i,\min}$ was zero. In cases where both r_i and G_i are zero, we obtained no information about the rate constant at all (see data in Figure 1c). Note that G_i is zero indicating that the reaction cannot occur at all during the corresponding time interval and, therefore, implies that $r_i=0$, but the converse does not hold.

To obtain good estimates, our aim is to ensure large values of both $r_{i,\min}$ and G_i by designing the experiment (i.e., manipulating the times at which measurements are taken). The guidelines we provide here are specific to stochastic chemical kinetic models and are not applicable during the exploratory phase of research when no reliable information is available about the system being measured. These guidelines are useful in the later stages of experimentation when we have a set of candidate reaction models in mind.

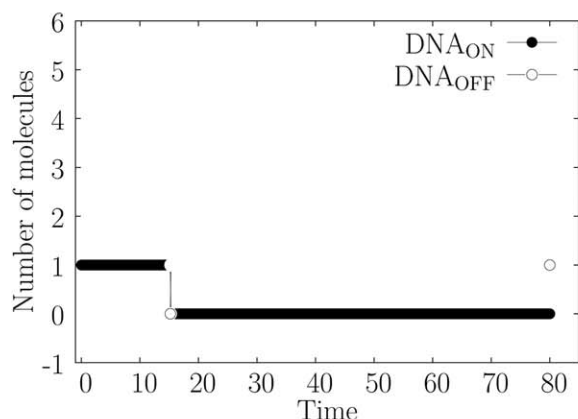
(1) As the parameter estimation depends on G_i (Eq. 14), it is more important to measure the reactants than the products. If the reactants are not measured, then the numbers of reactant molecules have to be inferred using Eq. 1 and statistical sampling, thereby increasing the computational burden, (2) measuring a reactant species that remains at zero does not provide information about the corresponding reaction(s). As we have seen in the early viral gene expression example, the mRNA (M) remains at zero in Figures 1e, f. The corresponding estimates of translation reaction rate constant k_f were close to zero. The measurement effort spent during the interval $[0, 10]$ when M was at zero, could have been better spent later during the experiment when the M was large enough. In some experiments, the measurements are left-censored, that is, we cannot measure them accurately until the measurement is above a certain threshold value. Obviously, measuring left-censored values provides little information, which is especially true for a stochastic chemical kinetic model, (3) estimation of a rate constant is difficult when we do not observe the occurrence of the corresponding reaction event. We have shown in the gene on-off example that estimates of k_{-1} were close to zero until we were able to observe \mathcal{R}_2 occurring, that is, until $r_{2,\min} > 0$. The same is true for the data in Figure 1e. For the early viral gene expression example, we know from VSV-BHK biology that the translation occurs later in the experiment, which allows



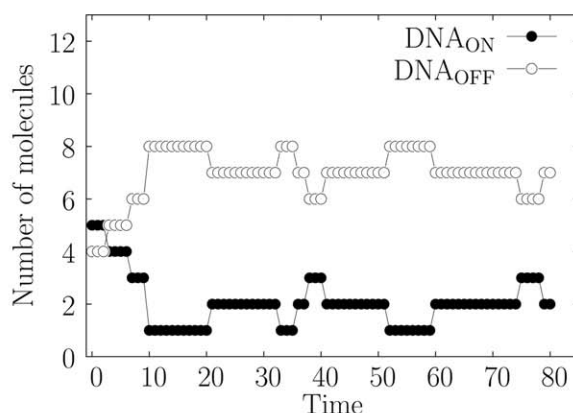
(a) $m = 4, \Delta s = 20$



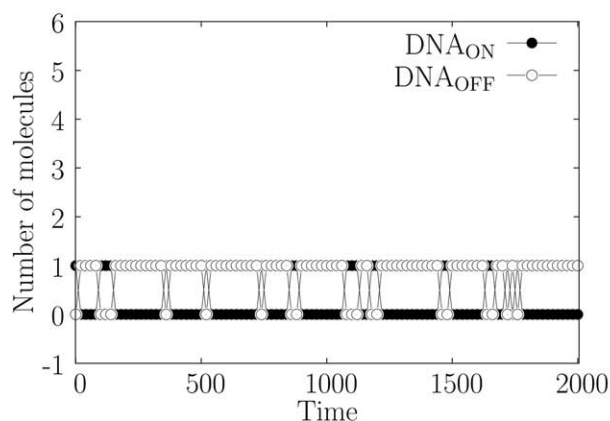
(b) $m = 4, \Delta s = 1$



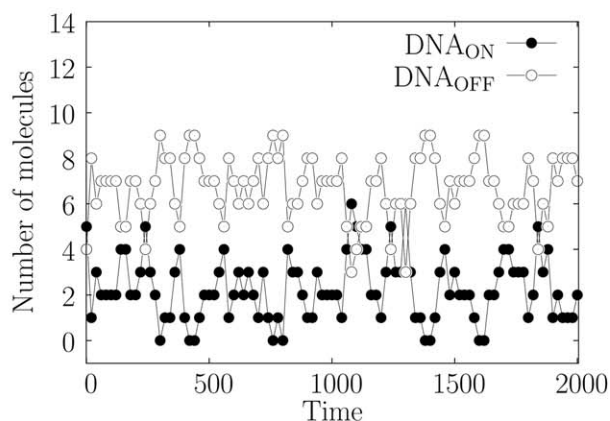
(c) $m = 800, \Delta s = 0.1$



(d) $m = 80, \Delta s = 1$



(e) $m = 100, \Delta s = 20$



(f) $m = 100, \Delta s = 20$

Figure 3. Gene on-off model.

True parameter values of $\theta_0 = [k_1, k_{-1}]^T = [0.03, 0.01]^T$. First column: $X(0) = [\text{DNA}_{\text{ON},0}, \text{DNA}_{\text{OFF},0}]^T = [1, 0]^T$. Second column: $X(0) = [\text{DNA}_{\text{ON},0}, \text{DNA}_{\text{OFF},0}]^T = [5, 4]^T$. m = number of points, Δs = sampling time.

us to reserve more resources for later measurements. For systems like the gene on-off model, the “gene switching on reaction” (\mathcal{R}_2) does not have such time dependence. Thus, the only way to observe the reaction events occur is by increasing the number of measurements. From the data in Figures 3d, f, we can see that it is better to perform the

same number measurements spread sparsely over a longer period of time instead of finely spread over a shorter period of time, and (4) increased measurement frequency does provide some benefit. Likelihood plots in Figures 4a, c show that increased measurement frequency does improve the estimate of k_1 . This increased measurement frequency does not

Table 4. Parameter Estimates and Sampling Time (Seconds per Sample) for Figure 3d

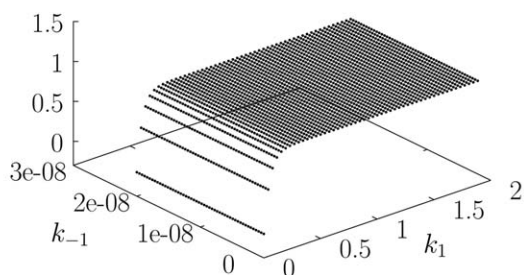
Parameter	Exact	Exact	CDIS	MCMC-MH	GibbsUnif
	(MLE)	(BE)	($N=100$)	($N=1000$)	($N=1000$)
k_1	0.0533	0.0534	0.0557	0.0556	11.9865
k_{-1}	0.0103	0.0103	0.0111	0.0115	3.5485
Sampling time (s/sample)	—	—	0.76	0.53	9.01

True values $\theta_0 = (k_{1,0}, k_{-1,0}) = (0.03, 0.01)$.

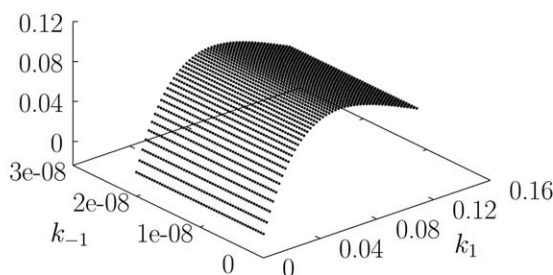
change the observed number of events, $r_{1,\min}$, but it does improve the value of G_1 , thus providing the improvement.

It is also important to choose the correct set of models for which to estimate parameters. If the model has a v^T with full column rank, then we can obtain the number of observed reaction events uniquely (i.e., $r_i = r_{i,\min}$) using the data and Eq. 1. If we can further identify that G_i is zero for a particular reaction (e.g., k_f in Figures 1e, f) then the data provides no information about the corresponding reaction, and we should

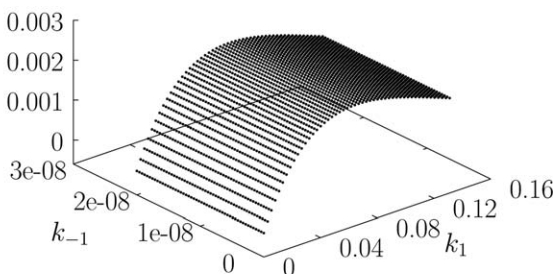
simplify the model by removing the reaction. Even if $G_i > 0$ but $r_i = 0$, then the estimate of the corresponding reaction rate constant is likely to be zero (e.g., k_r in Figures 1e, f) and this reaction may be removed as well. If the model's v^T does not have full column rank, the observed number of reaction events $r_{i,\min}$ provides similar guidelines. Using the exact likelihood method, we observed that estimates of k_{-1} were zero when $r_{2,\min} = 0$. In such cases, the “gene switching on” reaction \mathcal{R}_2 may be removed to simplify the model.



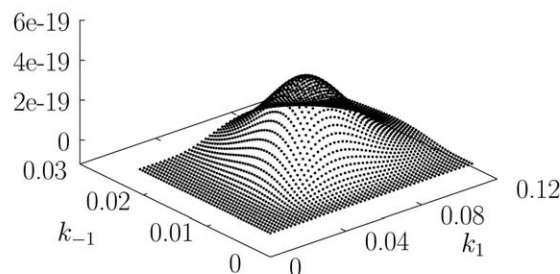
(a) $m = 4, \Delta s = 20$



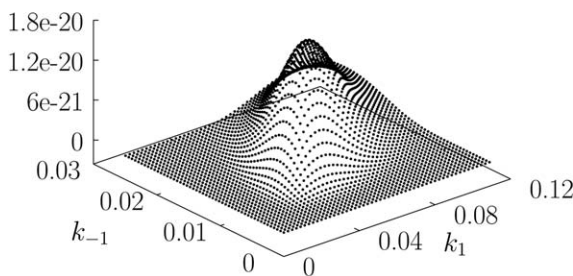
(b) $m = 4, \Delta s = 1$



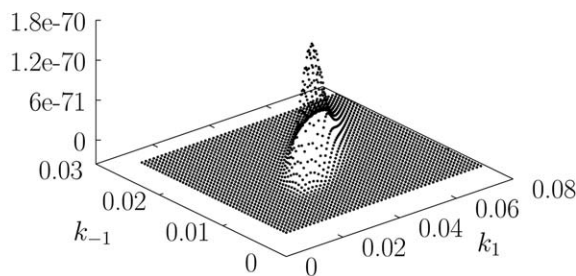
(c) $m = 800, \Delta s = 0.1$



(d) $m = 80, \Delta s = 1$



(e) $m = 100, \Delta s = 20$



(f) $m = 100, \Delta s = 20$

Figure 4. Gene on-off model. Exact likelihood plots for data in Figure 3.

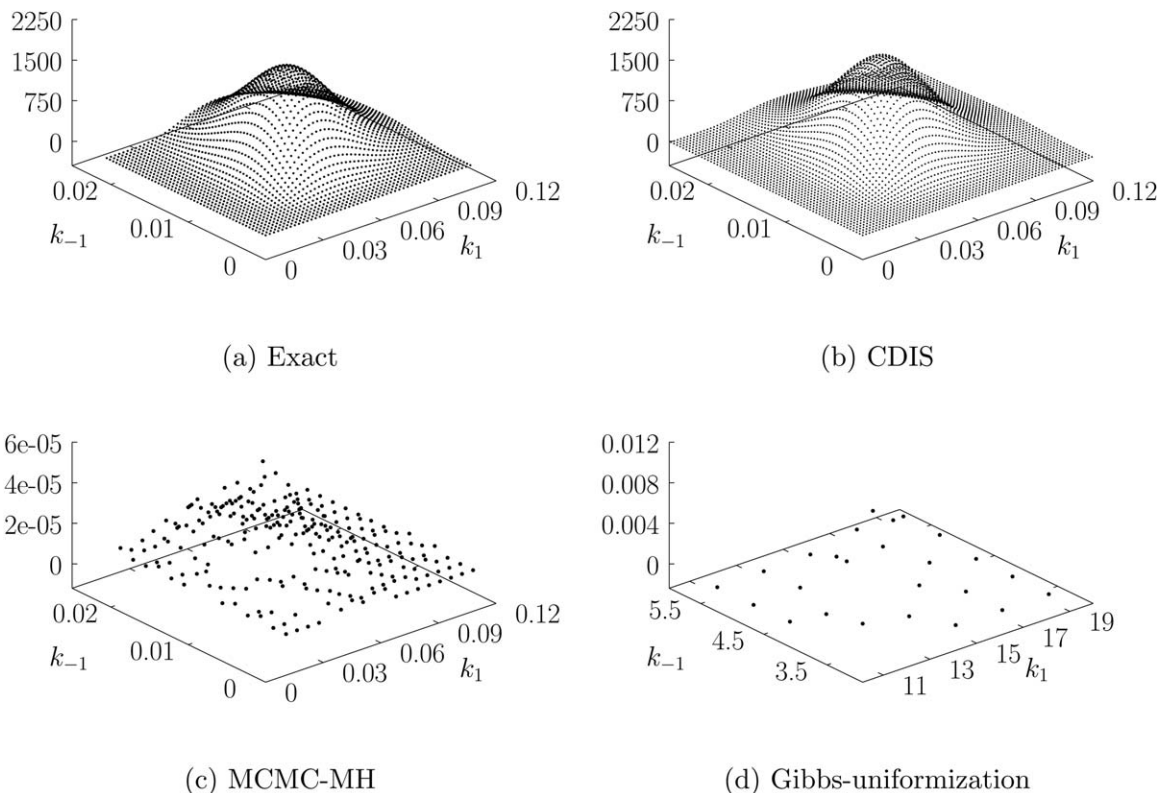


Figure 5. Joint posteriors for Figure 3d.

Conclusions and Future Work

In this article, we have demonstrated the use of three classes of parameter estimation methods, one of which is a new method. Using two examples from systems biology, we have demonstrated some strengths and weaknesses of all three methods. The exact method is intractable when the state space is large or unbounded (except in very special cases), which frequently happens in systems biology models. Even with a finite (and small) state space, Bayesian inference using the exact method requires numerical integration in n_r dimensions, which is also computationally expensive. The Gibbs-uniformization MCMC method, similar to the exact method, also uses the solution to Kolmogorov's forward equation and is not usually applicable. Furthermore, the uniformization technique requires enormous computation,

which makes the overall MCMC algorithm slow. More research is required to generate better endpoint-conditioned simulation methods. By comparison, the MCMC-MH method works well. All MCMC methods may sometimes have issues with convergence of Markov chains and require that a considerable fraction of Markov chain steps be burned-in, which increases the computational effort required. Even if these issues are resolved, MCMC methods provide the user with a histogram of the posterior that has the usual bias-variance tradeoff issues. The CDIS method, by contrast, provides a semianalytical expression for the posterior, which may be used for further analysis. We have shown that the CDIS method has good convergence properties and requires much less overall computation time than MCMC methods. Convergence of the CDIS method depends on the ability of the

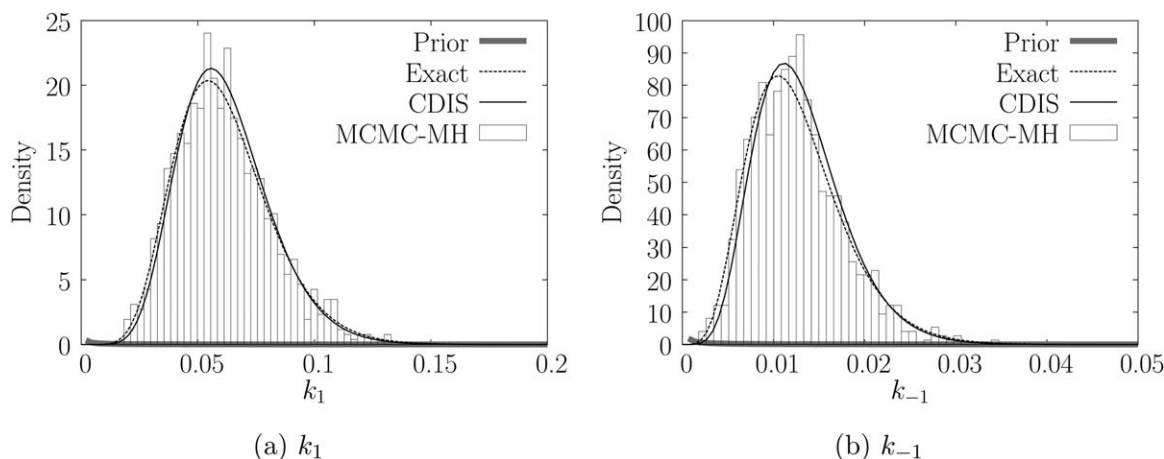


Figure 6. Marginal posteriors and priors for Figure 3d.

importance distribution to mimic the target distribution. There may be examples where the importance distribution we proposed is sufficiently different from the target distribution to require significantly more computation, but we have not yet encountered this issue.

Parameter estimation in stochastic chemical kinetics is a new area of research when compared to deterministic kinetics. Estimating parameters for a wide range of models using experimental data is a challenge that has two major bottlenecks. The first bottleneck is the measurements—only a partial set of species (usually one) can be measured at a time and even these measurements suffer from unknown measurement noise. The case of partial measurements is similar to the case of cyclical kinetics and may be solved by both CDIS and MCMC-MH methods. With the advancements in experimental techniques, the measurements are becoming more accurate and eventually it may be possible to obtain full measurements with low or at least characterized measurement noise. The second bottleneck is the computation time required. More measurements usually require more computation. Even with a small amount of “perfect data,” the computation time required may be large. In the future, when large amounts of imperfect data are available, computation is expected to be prohibitive, and fast, albeit approximate, estimators will be required.

Acknowledgments

The authors gratefully acknowledge funding for this research from a University of Wisconsin WARF Named Professorship Award. Andrea Timm and John Yin provided helpful comments and discussions. We are also grateful for support from the National Institutes of Health (AI071197 and AI091646)

Literature Cited

- McQuarrie DA. Stochastic approach to chemical kinetics. *J Appl Probab.* 1967;4:413–478.
- Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys.* 1976;22:403–434.
- Zheng Q, Ross J. Comparison of deterministic and stochastic kinetics for nonlinear systems. *J Chem Phys.* 1991;94:3644–3648.
- van Kampen NG. *Stochastic Processes in Physics and Chemistry*, 2nd ed. Amsterdam, The Netherlands: Elsevier Science Publishers, 1992.
- Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem.* 1977;81:2340–2361.
- Gibson MA, Bruck J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J Phys Chem A.* 2000;104:1876–1889.
- Anderson DF. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *J Chem Phys.* 2007;127:214107.
- Srivastava R, Haseltine EL, Mastny E, Rawlings JB. The stochastic quasi-steady-state assumption: reducing the model but not the noise. *J Chem Phys.* 2011;134:154109.
- Mastny EA, Haseltine EL, Rawlings JB. Two classes of quasi-steady-state model reductions for stochastic kinetics. *J Chem Phys.* 2007;127:094106.
- Cao Y, Gillespie DT, Petzold LR. The slow-scale stochastic simulation algorithm. *J Chem Phys.* 2005;122:014116.
- Haseltine EL, Rawlings JB. On the origins of approximations for stochastic chemical kinetics. *J Chem Phys.* 2005;123:164115.
- Haseltine EL, Rawlings JB. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J Chem Phys.* 2002;117:6959–6969.
- Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. *Science.* 2002;297:1183–1186.
- Blake WJ, Kaern MA, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. *Nature.* 2003;422:633–637.
- Raser JM, O’Shea EK. Control of stochasticity in eukaryotic gene expression. *Science.* 2004;304:1811–1814.
- McAdams HH, Arkin A. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA.* 1997;94:814–819.
- Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected *Escherichia coli* cells. *Genetics.* 1998;149:1633–1648.
- Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 tat fluctuations drive phenotypic diversity. *Cell.* 2005;122:169–182.
- Hensel SC, Rawlings J, Yin J. Stochastic kinetic modeling of Vesicular stomatitis virus intracellular growth. *Bull Math Biol.* 2009;71:1671–1692.
- Golding I, Paulsson J, Zawilski SM, Cox EC. Real-time kinetics of gene activity in individual bacteria. *Cell.* 2005;123:1025–1036.
- Tian T, Xu S, Gao J, Burrage K. Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics.* 2007;23:84–91.
- Poovathingal SK, Gunawan R. Global parameter estimation methods for stochastic biochemical systems. *BMC Bioinf.* 2010;11:1–12.
- Reinker S, Altman RM, Timmer J. Parameter estimation in stochastic biochemical reactions. *IEE Proc Syst Biol.* 2006;153:168–178.
- Wang Y, Christley S, Mjolsness E, Xie X. Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. *BMC Syst Biol.* 2010;4:1–16.
- Rempala GA, Ramos KS, Kalbfleisch T. A stochastic model of gene transcription: an application to L1 retrotransposition events. *J Theor Biol.* 2006;242:101–116.
- Henderson DA, Boys RJ, Krishnan KJ, Lawless C, Wilkinson DJ. Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons. *J Am Stat Assoc.* 2009;104:76–87.
- Golightly A, Wilkinson DJ. Bayesian sequential inference for stochastic kinetic biochemical network models. *J Comput Biol.* 2006;13:838–851.
- Golightly A, Wilkinson D. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Comput Stat Data Anal.* 2008;52:1674–1693.
- Golightly A, Wilkinson DJ. Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus.* 2011;1:807–820.
- Wilkinson DJ. *Stochastic Modelling for Systems Biology*, Vol. 44. Boca Raton, FL: CRC press, 2012.
- Boys RJ, Wilkinson DJ, Kirkwood TBL. Bayesian inference for a discretely observed stochastic kinetic model. *Stat Comput.* 2008;18:125–135.
- Choi B, Rempala GA. Inference for discretely observed stochastic kinetic networks with applications to epidemic modeling. *Biostatistics.* 2012;13:153–165.
- Hobolth A, Stone EA. Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution. *Ann Appl Stat.* 2009;3:1204–1231.
- Rawlings JB, Ekerdt JG. *Chemical Reactor Analysis and Design Fundamentals*. Madison, WI: Nob Hill Publishing, 2004:640 p.
- Chen W-Y, Kulkarni A, Milum JL, Fan LT. Stochastic modeling of carbon oxidation. *AIChE J.* 1999;45:2557–2570.
- Graham MD, Rawlings JB. *Modeling and Analysis Principles for Chemical and Biological Engineers*. Madison, WI: Nob Hill Publishing, 2013:552 p.
- Jahnke T, Huisinga W. Solving the chemical master equation for monomolecular reaction systems analytically. *J Math Biol.* 2007;54:1–26.
- Asmussen S, Hobolth A. Bisection ideas in end-point conditioned Markov process simulation. In: *Proceedings of the 7th International Workshop on Rare Event Simulation*. Rennes, France, Vol. 69. 2008:7499–7506.
- Gupta A. Parameter estimation in deterministic and stochastic models of biological systems, Ph.D. thesis, University of Wisconsin-Madison, Madison, WI, 2013.
- Timm A, Yin J. Kinetics of virus production from single cells. *Virology.* 2012;424:11–17.
- Rose JK, Whitt MA. *Rhabdoviridae: The viruses and their replication*. In: Knipe DM, Howley PM, editor. *Fields Virology*, Vol. 1, 4th ed. Philadelphia: Lippincott Williams & Wilkins, 2001:1221–1244.
- Wasserman L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

Manuscript received Oct. 26, 2013, and final revision received Feb. 11, 2014.