Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems

John Goutsias

Citation: J. Chem. Phys. **122**, 184102 (2005); doi: 10.1063/1.1889434
View online: https://doi.org/10.1063/1.1889434
View Table of Contents: http://aip.scitation.org/toc/jcp/122/18
Published by the American Institute of Physics

Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems

John Goutsias[a)]
The Whitaker Biomedical Engineering Institute, Johns Hopkins University, Baltimore, Maryland 21218

(Received 18 November 2004; accepted 18 February 2005; published online 9 May 2005)

We address the problem of eliminating fast reaction kinetics in stochastic biochemical systems by employing a quasiequilibrium approximation. We build on two previous methodologies developed by [Haseltine and Rawlings, J. Chem. Phys. **117**, 6959 (2002)] and by [Rao and Arkin, J. Chem. Phys. **118**, 4999 (2003)]. By following Haseltine and Rawlings, we use the numbers of occurrences of the underlying reactions to characterize the state of a biochemical system. We consider systems that can be effectively partitioned into two distinct subsystems, one that comprises "slow" reactions and one that comprises "fast" reactions. We show that when the probabilities of occurrence of the slow reactions depend at most *linearly* on the states of the fast reactions, we can effectively eliminate the fast reactions by modifying the probabilities of occurrence of the slow reactions. This modification requires computation of the mean states of the fast reactions, conditioned on the states of the slow reactions. By assuming that within consecutive occurrences of slow reactions, the fast reactions rapidly reach equilibrium, we show that the conditional state means of the fast reactions satisfy a system of at most *quadratic* equations, subject to *linear* inequality constraints. We present three examples which allow analytical calculations that clearly illustrate the mathematical steps underlying the proposed approximation and demonstrate the accuracy and effectiveness of our method. © *2005 American Institute of Physics*. [DOI: 10.1063/1.1889434]

I. INTRODUCTION

Modeling interactions among molecular species in single cells and simulating the dynamic behavior determined by such interactions are two very important problems of computational cell biology. In single cells, molecular interactions produce stochastic fluctuations that may dominate cellular dynamics.[1,2] Therefore, modeling molecular interactions in single cells requires use of *stochastic* biochemical systems, which provide a probabilistic description of cellular dynamics.

Modeling and simulation of a highly reactive stochastic biochemical system comprising many species is difficult, since each molecule and every reaction event should be accounted for. Gillespie has shown that a well-mixed biochemical system at thermal equilibrium can be effectively characterized by a time-dependent joint probability mass function (PMF) over the system state.[3] A first-order differential equation, known as the chemical master equation (CME), governs the time evolution of this PMF. Unfortunately, modeling a complex biochemical system most often leads to a *nonlinear* CME, which cannot be solved analytically. However, it is possible to simulate the dynamics of the system by a stochastic algorithm known as the Gillespie algorithm[4] and estimate relevant statistics (e.g., means, variances, and PMFs) by Monte Carlo simulation.[5]

Although the Gillespie algorithm is a method for *exact* simulation, it is often computationally intensive. This is especially true in cases of highly reactive biochemical systems comprising a large number of molecular species. If the biochemical system is highly reactive, reactions will occur numerous times in a short period of time and the Gillespie algorithm will be spending most time to simulate these events. Attempts to accelerate the Gillespie algorithm have produced several refinements.[6–13] However, some of these refinements are not appropriate for simulating biochemical systems in which the populations of all species are small, whereas many of these refinements remain computationally expensive as biochemical systems become progressively more complex.

Recently, Rao and Arkin[14] proposed a technique to reduce system complexity by partitioning molecular species into two groups, namely, into primary species and intermediate species. By introducing an appropriate time-scale separation, they assume that the intermediate species are asymptotically at steady state within a coarse time scale. Then, they derive a CME that approximately characterizes the stochastic evolution of the primary species, thus effectively approximating the biochemical system by eliminating all intermediate species. The intermediate species exert their influence on the coefficients of the approximating CME, which are modified to account for the elimination of these species.

The method proposed by Rao and Arkin relies on two key assumptions: (a) given the population numbers of the primary species, the stochastic evolution of the population numbers of the intermediate species is Markovian and (b)

a)Electronic mail: goutsias@jhu.edu

given the population numbers of the primary species, the *net* rate of change of the conditional joint PMF of the population numbers of the intermediate species is ≈ 0. Although the second assumption derives naturally from a statistical interpretation of the well-known quasisteady-state assumption (QSSA) of deterministic chemical kinetics,[15] the first assumption has not been theoretically justified. Therefore, it is not clear in which cases and under what conditions this assumption is valid. Moreover, the QSSA is not always appropriate and may lead to erroneous approximations. This is illustrated by a simple enzyme kinetics example presented by Rao and Arkin.[14]

An attractive approach to reduce system complexity is to recognize that, in certain cases, biochemical reactions may be separated into "slow" and "fast" reactions. Slow reactions occur infrequently over a long period of time, whereas fast reactions occur numerous times over a short period of time. It is the presence of fast reactions which dramatically increases computations. However, fast reactions are usually transitory and of no particular interest. Therefore, elimination of these reactions may result in a simpler biochemical system, which can be modeled easier and simulated faster than the original system.

An approach for eliminating fast reaction kinetics has been proposed by Haseltine and Rawlings.[16] In their approach, the system state is specified by the numbers of occurrences of the underlying reactions. By approximating the fast reactions with a Langevin equation, they propose a CME for characterizing the stochastic evolution of the slow reactions. Although this approach works well in a simple crystallization and a viral infection example,[16] it suffers from several drawbacks. For example, the coefficients of the approximating CME may substantially vary within consecutive occurrences of slow reactions. This requires use of a modified version of the Gillespie algorithm that explicitly deals with this problem. The modified algorithm is more difficult to use than the standard Gillespie algorithm and may result in a computationally expensive implementation. To ameliorate this problem, Haseltine and Rawlings proposed an additional approximation, based on scaling the stochastic step of the Gillespie algorithm. However, it is not clear how this approximation affects the accuracy and computational complexity of the proposed algorithm. Moreover, approximating a stochastic biochemical system with the CME proposed by Haseltine and Rawlings may not be theoretically sound, since we rigorously show in this paper that the stochastic evolution of the slow reactions is approximately governed by another CME, which does not necessarily imply the CME proposed by Haseltine and Rawlings (see our discussion in Sec. III).

In this paper, we introduce a new approach to the problem of eliminating fast reaction kinetics, which leads to a method that avoids the previous problems. Initially, our approach follows that of Haseltine and Rawlings. However, after a brief introduction of the CME in Sec. II, we rigorously derive a CME in Sec. III which effectively models the dynamic evolution of the marginal PMF of a biochemical system over its slow states. If the coefficients of this marginal CME are known, then it can be stochastically simulated by the standard Gillespie algorithm. We show that for stochastic biochemical systems in which the probability of occurrence of a slow reaction depends at most *linearly* on the numbers of occurrences of the fast reactions, the coefficients of the marginal CME can be specified from the coefficients of the original CME and knowledge of the mean states of the fast reactions, conditioned on the state values of the slow reactions. To illustrate the proposed method, we present three examples: an enzyme kinetics example in Sec. IV, a viral infection kinetics example in Sec. V, and a transcription regulation example in Sec. VI. It becomes clear from these examples that, by assuming the fast reactions to be at equilibrium within consecutive occurrences of slow reactions, the conditional state means of the fast reactions satisfy a system of at most *quadratic* equations, subject to linear inequality constraints. Therefore, specification of the marginal CME requires solution of a constrained system of (at most quadratic) equations. In all three examples, this can be done analytically. The examples clearly illustrate the accuracy and effectiveness of the proposed approximation technique. Finally, we conclude the paper in Sec. VII by discussing several issues, including a possible numerical procedure for calculating the coefficients of the marginal CME when analytical derivation of such coefficients is not possible.

II. THE CHEMICAL MASTER EQUATION

Consider a well-mixed biochemical system at thermal equilibrium consisting of M *elementary* (monomolecular or bimolecular) *irreversible* reaction channels. We model a reversible reaction by two separate irreversible reactions and a reaction that involves more than two molecules as a cascade of bimolecular reactions.[17] We may characterize the state of the biochemical system at time $t \geqslant 0$ by the N-dimensional random vector $\mathbf{X}(t)=[X_1(t)X_2(t)\cdots X_N(t)]'$, where $X_n(t)$ is the number of molecules of the nth reactant or product species present in the system at time t and $\mathbf{X}'$ denotes transposition. Given that the biochemical system is at state $\mathbf{X}(t)=\mathbf{x}$ at time t, the probability that one mth reaction will occur during the time interval $[t,t+dt)$ is approximately given by $\pi_m(\mathbf{x})dt$ for a sufficiently small dt, where $\pi_m(\mathbf{x})$ is the *propensity function* of the mth reaction channel.[3,18] This function is expressed as

$$\pi_m(\mathbf{x}) = c_m h_m(\mathbf{x}), \quad m \in \mathcal{M}, \tag{1}$$

where $\mathcal{M}=\{1,2,\ldots,M\}$, $c_m>0$ is the *specific probability rate constant* of the mth reaction, and $h_m(\mathbf{x})$ is the number of all possible distinct combinations of the reactant molecules associated with the mth reaction channel when the system is at state $\mathbf{x}$, given by

$$h_m(\mathbf{x}) = \begin{cases} x_i & \text{for monomolecular reactions} \\ x_i(x_i-1)/2 & \text{for bimolecular reactions with identical reactants} \\ x_i x_j & \text{for bimolecular reactions with different reactants,} \end{cases} \quad (2)$$

for some $1 \leqslant i, j \leqslant N$, $i \neq j$. The specific probability rate constant c_m is the probability per unit time that a randomly chosen combination of reactant molecules of the mth reaction will react. If the reaction rate constant k_m of the mth reaction is known, then[3,18]

$$c_m = \begin{cases} k_m & \text{for monomolecular reactions} \\ 2k_m/AV & \text{for bimolecular reactions with identical reactants} \\ k_m/AV & \text{for bimolecular reactions with distinct reactants,} \end{cases} \quad (3)$$

where $A = 6.022\,141\,5 \times 10^{23}$ mol^{-1} is the Avogadro constant and V is the cellular volume (in liters). From Eqs. (1) and (2), note that the propensities are linear or quadratic functions of $\mathbf{x}$.

An alternative way to characterize the state of a biochemical system at time $t \geqslant 0$ is by means of the M-dimensional random vector $\mathbf{Z}(t) = [Z_1(t) Z_2(t) \cdots Z_M(t)]'$, where $Z_m(t) = z \geqslant 0$, if the mth reaction has occurred z times during the time interval $[0, t)$. We refer to the random variable $Z_m(t)$ as the *degree of advancement* (DA) of the mth reaction.[19,20] $\mathbf{X}(t)$ can be uniquely determined from $\mathbf{Z}(t)$, since

$$\mathbf{X}(t) = \mathbf{x}_0 + \mathbb{S}\mathbf{Z}(t), \quad t \geqslant 0, \quad (4)$$

where $\mathbf{x}_0$ is the initial molecular state, $\mathbb{S}$ is the $N \times M$ *stoichiometry matrix* of the biochemical system, and $\mathbf{Z}(0) = \mathbf{0}$, where $\mathbf{0}$ is the null vector. Note that $\mathbb{S} = [\mathbf{s}_1 \mathbf{s}_2 \cdots \mathbf{s}_M]$, where $\mathbf{s}_m = [s_{1m} s_{2m} \cdots s_{Nm}]'$ is the N-dimensional vector of the *stoichiometric coefficients* associated with the mth reaction. The stoichiometric coefficient s_{nm} is the change in the number of molecules of the nth species caused by one occurrence of the mth reaction.[15]

To characterize the discrete-valued DA process $\mathbf{Z} = \{\mathbf{Z}(t), t \geqslant 0\}$, we need to specify the joint PMF $P(\mathbf{z}; t) = \Pr[\mathbf{Z}(t) = \mathbf{z} | \mathbf{Z}(0) = \mathbf{0}]$, for every $t \geqslant 0$. By following Gillespie,[3] we can show that $P(\mathbf{z}; t)$ satisfies the following CME (see also Haseltine and Rawlings[16]):

$$\frac{\partial P(\mathbf{z}; t)}{\partial t} = \sum_{m \in \mathcal{M}} \alpha_m(\mathbf{z} - \mathbf{e}_m) P(\mathbf{z} - \mathbf{e}_m; t) - \alpha_m(\mathbf{z}) P(\mathbf{z}; t), \quad (5)$$

with initial condition $P(\mathbf{0}; 0) = 1$, where

$$\alpha_m(\mathbf{z}) \triangleq \pi_m(\mathbf{x}_0 + \mathbb{S}\mathbf{z}) = c_m h_m(\mathbf{x}_0 + \mathbb{S}\mathbf{z}), \quad (6)$$

and $\mathbf{e}_m$ is the mth column of the $M \times M$ identity matrix. Equation (5) implies that $\mathbf{Z}$ is a multivariate Markov process. As a matter of fact, $\mathbf{Z}$ is a multivariate *birth* process.[21]

We would like to point out that, in contrast to our approach in this paper, it is most common in the literature to characterize the state of a stochastic biochemical system by the population process $\mathbf{X} = \{\mathbf{X}(t), t \geqslant 0\}$ and not by the DA process $\mathbf{Z}$. Note that, although we can use Eq. (4) to uniquely determine $\mathbf{X}$ from $\mathbf{Z}$, we cannot in general determine $\mathbf{Z}$ from $\mathbf{X}$. This is due to the fact that there might be two different sets of reactions that may lead to the same molecular populations. We can show that, when $\mathbf{X}$ is given by Eq. (4), its joint PMF satisfies a CME similar to Eq. (5).[3,14] However, many important biochemical systems, such as systems of genetic regulation, are subject to *nonzero* delays between the times when reactions are initiated and the times when products become available. In this case, we may assume that each reaction m is characterized by a fixed time delay τ_m and set [compare with Eq. (4)]

$$X_n(t) = x_{0,n} + \sum_{m \in \mathcal{M}} s_{nm} Z_m(t - \tau_m), \; t \geqslant 0, \; n = 1, 2, \ldots, N. \quad (7)$$

When the population process $\mathbf{X}$ is given by Eq. (7), we cannot in general derive a CME for its joint PMF and, therefore, we cannot use $\mathbf{X}$ to describe the biochemical system. However, we can model the state of the system by the DA process $\mathbf{Z}$ and use Eq. (7) to obtain $\mathbf{X}$ from $\mathbf{Z}$. Therefore, the DA process $\mathbf{Z}$ provides a more appropriate characterization of a stochastic biochemical system than the population process $\mathbf{X}$.

III. PROBABILISTIC SEPARATION OF REACTION KINETICS

Several steps, similar to the ones followed by Haseltine and Rawlings[16] and by Rao and Arkin,[14] can be used to probabilistically separate the slow reaction channels from the fast reaction channels in a stochastic biochemical system. To distinguish between slow and fast reaction channels, we need to determine an appropriate *coarse* time scale over which a group of channels rarely react, whereas the remaining channels react very frequently. It is expected that, over the coarse time scale, the DAs associated with the first group of reactions (the slow reactions) will almost never change value, whereas the DAs associated with the second group of reactions (the fast reactions) will change values numerous times. We assume that the first M_0 reactions of a biochemical system are slow, whereas the remaining $M - M_0$ reactions are fast, and set

$$\mathbf{Z}(t) = \begin{bmatrix} \mathbf{Z}_s(t) \\ \mathbf{Z}_f(t) \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} \mathbf{z}_s \\ \mathbf{z}_f \end{bmatrix},$$

$$\mathbf{e}_m = \begin{bmatrix} \overline{\mathbf{e}}_m \\ \mathbf{0} \end{bmatrix}, \quad m \in \mathcal{M}_s,$$

and

$$\mathbf{e}_m = \begin{bmatrix} \mathbf{0} \\ \underline{\mathbf{e}}_m \end{bmatrix}, \quad m \in \mathcal{M}_f,$$

where $\mathcal{M}_s = \{1, 2, \ldots, M_0\}$ and $\mathcal{M}_f = \{M_0+1, M_0+2, \ldots, M\}$. In the previous equations, $\mathbf{Z}_s(t)$, $\mathbf{z}_s$, and $\overline{\mathbf{e}}_m$ are M_0-dimensional vectors, whereas $\mathbf{Z}_f(t)$, $\mathbf{z}_f$, and $\underline{\mathbf{e}}_m$ are $(M - M_0)$-dimensional vectors. Then, by summing both sides of Eq. (5) with respect to $\mathbf{z}_f$ and by using the fact that

$$P(\mathbf{z}_s, \mathbf{z}_f; t) = P(\mathbf{z}_f | \mathbf{z}_s; t) P(\mathbf{z}_s; t), \tag{8}$$

we obtain

$$\frac{\partial P(\mathbf{z}_s; t)}{\partial t} = \sum_{m \in \mathcal{M}_s} \alpha_m^{(t)}(\mathbf{z}_s - \overline{\mathbf{e}}_m) P(\mathbf{z}_s - \overline{\mathbf{e}}_m; t) - \alpha_m^{(t)}(\mathbf{z}_s) P(\mathbf{z}_s; t), \tag{9}$$

where

$$\alpha_m^{(t)}(\mathbf{z}_s) \triangleq \sum_{\mathbf{z}_f} \alpha_m(\mathbf{z}_s, \mathbf{z}_f) P(\mathbf{z}_f | \mathbf{z}_s; t), \quad m \in \mathcal{M}_s. \tag{10}$$

Equation (9) shows that the dynamic evolution of the marginal PMF $P(\mathbf{z}_s; t)$ of the slow reaction subsystem is governed by a CME, similar to the CME (5), which includes only the slow reaction channels. The fast reactions exert their influence on the slow reaction subsystem through the propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, where $\alpha_m^{(t)}(\mathbf{z}_s)$ is the conditional mean of the propensity function $\alpha_m(\mathbf{z}_s, \mathbf{z}_f)$ of the mth slow reaction channel with respect to the fast DAs $\mathbf{Z}_f(t)$, conditioned on the values $\mathbf{z}_s(t)$ of the slow DAs. Note that Eqs. (9) and (10) are the analogs of Eqs. (10) and (11) of Rao and Arkin.[14] The Rao and Arkin equations are based on separating the biochemical system into primary and intermediate species, whereas our equations are based on a more natural separation of the system into slow and fast reactions. Furthermore, the Rao and Arkin equations govern the dynamic evolution of the joint PMF of the population numbers of the primary species. Our equations govern the dynamic evolution of the joint PMF of the slow DAs, which is advantageous when modeling biochemical systems with nonzero delays, as we explained in Sec. II. Finally, the Rao and Arkin equations are based on the assumptions that, given the population numbers of the primary species, the stochastic evolution of the population numbers of the intermediate species is Markovian and that the *net* rate of change of the conditional joint PMF of the population numbers of the intermediate species is ≈ 0. However, the derivation of Eqs. (9) and (10) requires no assumptions. These equations are exact.

To utilize the CME (9), we need to know the propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, whose calculation requires knowledge of the conditional PMF $P(\mathbf{z}_f | \mathbf{z}_s; t)$. We can show that, within the coarse time scale, the dynamic evolution of the conditional PMF $P(\mathbf{z}_f | \mathbf{z}_s; t)$ of the fast reaction subsystem, given the state of the slow reaction subsystem, is *approximately* governed by the CME

$$\frac{\partial P(\mathbf{z}_f | \mathbf{z}_s; t)}{\partial t} = \sum_{m \in \mathcal{M}_f} \alpha_m(\mathbf{z}_s, \mathbf{z}_f - \underline{\mathbf{e}}_m) P(\mathbf{z}_f - \underline{\mathbf{e}}_m | \mathbf{z}_s; t) - \alpha_m(\mathbf{z}_s, \mathbf{z}_f) P(\mathbf{z}_f | \mathbf{z}_s; t), \tag{11}$$

which includes only the fast reaction channels.

Indeed, since it is very unlikely that a slow reaction will occur within the coarse time scale, we may approximately assume that, within that time scale, the biochemical system behaves like one with slow propensity functions that are not appreciably larger than zero. Therefore, we may set $\alpha_m(\mathbf{z}_s, \mathbf{z}_f) \simeq 0$ for $m \in \mathcal{M}_s$. In this case, Eq. (5) becomes

$$\frac{\partial P(\mathbf{z}_s, \mathbf{z}_f; t)}{\partial t} \simeq \sum_{m \in \mathcal{M}_f} \alpha_m(\mathbf{z}_s, \mathbf{z}_f - \underline{\mathbf{e}}_m) P(\mathbf{z}_s, \mathbf{z}_f - \underline{\mathbf{e}}_m; t) - \alpha_m(\mathbf{z}_s, \mathbf{z}_f) P(\mathbf{z}_s, \mathbf{z}_f; t), \tag{12}$$

whereas Eq. (9) results in

$$\frac{\partial P(\mathbf{z}_s; t)}{\partial t} \simeq 0. \tag{13}$$

However, from Eq. (8), we have that

$$\frac{\partial P(\mathbf{z}_s, \mathbf{z}_f; t)}{\partial t} = P(\mathbf{z}_f | \mathbf{z}_s; t) \frac{\partial P(\mathbf{z}_s; t)}{\partial t} + \frac{\partial P(\mathbf{z}_f | \mathbf{z}_s; t)}{\partial t} P(\mathbf{z}_s; t),$$

which, together with Eq. (13), results in

$$\frac{\partial P(\mathbf{z}_s, \mathbf{z}_f; t)}{\partial t} \simeq \frac{\partial P(\mathbf{z}_f | \mathbf{z}_s; t)}{\partial t} P(\mathbf{z}_s; t). \tag{14}$$

Finally, Eq. (11) is approximately obtained from Eqs. (8), (12), and (14).

Note that Eq. (11) is the analog of Eq. (5) of Rao and Arkin,[14] which governs the dynamic evolution of the joint PMF of the population numbers of the intermediate species, given the population numbers of the primary species. However, Eq. (5) of Rao and Arkin has been derived by assuming that the "intermediate" state is Markovian, given the "primary" state. We have shown that Eq. (11) can be derived by naturally assuming that, within the coarse time scale, the biochemical system behaves like one whose slow propensity functions are ≈ 0.

Unfortunately, solving Eq. (11) is not possible in general. Moreover, calculating the propensity function $\alpha_m^{(t)}(\mathbf{z}_s)$, for $m \in \mathcal{M}_s$, requires evaluation of a summation over all possible states $\mathbf{z}_f$. In principle, this summation can be estimated by Monte Carlo simulation based on samples $\mathbf{z}_f$ drawn from the CME (11) using the Gillespie algorithm. However, this is a computationally daunting task. Therefore, calculation of the propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, by means of Eq. (10), is not possible in general.

If we limit our interest to stochastic biochemical systems for which the slow propensity functions depend linearly on fast DAs, we can show that

$$\alpha_m^{(t)}(\mathbf{z}_s) = \alpha_m[\mathbf{z}_s, \mu(\mathbf{z}_s; t)], \quad m \in \mathcal{M}_s, \tag{15}$$

where

$$\mu(\mathbf{z}_s;t) \triangleq [\mu_{M_0+1}(\mathbf{z}_s;t)\mu_{M_0+2}(\mathbf{z}_s;t)\cdots\mu_M(\mathbf{z}_s;t)]',$$

with $\mu_m(\mathbf{z}_s;t)$ being the mean DA of the mth fast reaction at time t, given the state $\mathbf{z}_s$ of the slow reaction subsystem at t; i.e.,

$$\mu_m(\mathbf{z}_s;t) \triangleq \mathrm{E}[Z_m(t)|\mathbf{Z}_s(t)=\mathbf{z}_s], \quad m \in \mathcal{M}_f. \tag{16}$$

Indeed, a Taylor series expansion of the slow propensity function $\alpha_m(\mathbf{z}_s,\mathbf{z}_f)$, $m \in \mathcal{M}_s$, with respect to $\mathbf{z}_f$ about the conditional mean vector $\mu(\mathbf{z}_s;t)$ gives

$$\begin{aligned}\alpha_m(\mathbf{z}_s,\mathbf{Z}_f) &= \alpha_m(\mathbf{z}_s,\mu) \\ &\quad + (\mathbf{Z}_f-\mu)'\frac{\partial\alpha_m(\mathbf{z}_s,\mu)}{\partial\mathbf{z}_f} \\ &\quad + \frac{1}{2}(\mathbf{Z}_f-\mu)'\frac{\partial^2\alpha_m(\mathbf{z}_s,\mu)}{\partial\mathbf{z}_f^2}(\mathbf{Z}_f-\mu)\end{aligned} \tag{17}$$

for $m \in \mathcal{M}_s$, where $\partial g(\mathbf{z})/\partial\mathbf{z}$ denotes a vector with elements $\partial g(\mathbf{z})/\partial z_m$ and $\partial^2 g(\mathbf{z})/\partial\mathbf{z}^2$ denotes a matrix with elements $\partial^2 g(\mathbf{z})/\partial z_m\partial z_n$. From Eqs. (2) and (6), note that the derivatives of $\alpha_m(\mathbf{z}_s,\mathbf{z}_f)$ with respect to $\mathbf{z}_f$ of order greater than 2 are zero, due to the linear or quadratic nature of $h_m(\mathbf{x})$. Therefore, all terms of the Taylor series expansion in Eq. (17) of order $\geqslant 3$ are zero. If the propensity functions of the slow bimolecular reactions depend linearly on fast DAs, then the third term on the right-hand side of Eq. (17) is zero, and

$$\alpha_m(\mathbf{z}_s,\mathbf{Z}_f) = \alpha_m(\mathbf{z}_s,\mu) + (\mathbf{Z}_f-\mu)'\frac{\partial\alpha_m(\mathbf{z}_s,\mu)}{\partial\mathbf{z}_f}, \quad m \in \mathcal{M}_s. \tag{18}$$

By setting $\mathbf{Z}_f \to \mathbf{Z}_f(t)$ in Eq. (18) and by taking the mean on both sides of that equation with respect to $\mathbf{Z}_f(t)$, conditioned on $\mathbf{Z}_s(t)=\mathbf{z}_s$, we obtain Eq. (15).

Equation (15) shows that the fast reaction kinetics of a stochastic biochemical system with slow propensity functions that depend linearly on fast DAs can be effectively summarized by the conditional mean DAs $\mu_m(\mathbf{z}_s;t)$, $m \in \mathcal{M}_f$, associated with the fast reactions. Therefore, the problem of approximating fast reaction kinetics reduces to the problem of calculating the conditional mean DAs $\mu_m(\mathbf{z}_s;t)$, $m \in \mathcal{M}_f$, of the fast reaction subsystem. In the rest of the paper, we present three examples that show how to approximate these means and illustrate the accuracy and effectiveness of the proposed approach.

We conclude this section by noting that Haseltine and Rawlings approximate the slow reaction subsystem by the following CME:

$$\begin{aligned}\frac{\partial P(\mathbf{z}_s;t)}{\partial t} &= \sum_{m\in\mathcal{M}_s}\alpha_m[\mathbf{z}_s-\overline{\mathbf{e}}_m,\mathbf{Z}_f(t)]P(\mathbf{z}_s-\overline{\mathbf{e}}_m;t) \\ &\quad - \alpha_m[\mathbf{z}_s,\mathbf{Z}_f(t)]P(\mathbf{z}_s;t),\end{aligned} \tag{19}$$

where $\mathbf{Z}_f(t)$ satisfies an appropriate Langevin equation.[16] Note, however, that the validity of this CME is questionable, since we have rigorously shown above that the PMF $P(\mathbf{z}_s;t)$ satisfies the CME (9), which does not necessarily imply Eq. (19).

IV. EXAMPLE: ENZYME KINETICS

As a first example, we consider a common mechanism for enzyme catalyzed reactions, governed by the following three elementary reactions:[17]

$$\begin{aligned}&\text{reaction 1:} \quad \mathrm{E}\cdot\mathrm{S}\xrightarrow{c_1}\mathrm{P}+\mathrm{E}, \\ &\text{reaction 2:} \quad \mathrm{E}+\mathrm{S}\xrightarrow{c_2}\mathrm{E}\cdot\mathrm{S}, \\ &\text{reaction 3:} \quad \mathrm{E}\cdot S\xrightarrow{c_3}E+\mathrm{S}.\end{aligned} \tag{20}$$

The biochemical system under consideration consists of enzymes E, which catalyze the reaction of substrates S into products P by forming intermediate enzyme/substrate complexes $\mathrm{E}\cdot\mathrm{S}$. This simple example allows us to illustrate our approximation scheme and compare it to the technique proposed by Rao and Arkin.[14]

We may use variables $X_n(t)$, $n=1,2,3,4$, to characterize the molecular state of the biochemical system at time $t \geqslant 0$. Each variable denotes the number of molecules of a reactant or product species, as identified by the following assignment:

$$X_n \leftrightarrow \begin{cases}\mathrm{P} & \text{for } n=1 \\ \mathrm{E} & \text{for } n=2 \\ \mathrm{S} & \text{for } n=3 \\ \mathrm{E}\cdot\mathrm{S} & \text{for } n=4.\end{cases}$$

From Eqs. (2) and (20), we have that

$$\begin{aligned}h_1(x_1,x_2,x_3,x_4) &= x_4, \\ h_2(x_1,x_2,x_3,x_4) &= x_2x_3, \\ h_3(x_1,x_2,x_3,x_4) &= x_4,\end{aligned} \tag{21}$$

whereas the stoichiometry matrix is given by

$$\mathbb{S} = \begin{bmatrix}1 & 0 & 0 \\ 1 & -1 & 1 \\ 0 & -1 & 1 \\ -1 & 1 & -1\end{bmatrix}. \tag{22}$$

We initialize the system with $e \geqslant 1$ enzyme molecules and $s \geqslant 1$ substrate molecules, in which case

$$\mathbf{x}_0 = [0 \quad e \quad s \quad 0]'. \tag{23}$$

From Eqs. (4), (22), and (23), we have

$$\begin{aligned}X_1(t) &= Z_1(t), \\ X_2(t) &= e + Z_1(t) - Z_2(t) + Z_3(t), \\ X_3(t) &= s - Z_2(t) + Z_3(t), \\ X_4(t) &= -Z_1(t) + Z_2(t) - Z_3(t).\end{aligned} \tag{24}$$

Finally, note that the CME underlying the stochastic evolution of the associated DAs is given by Eq. (5), with propensity functions [see Eqs. (6), (21), and (24)]

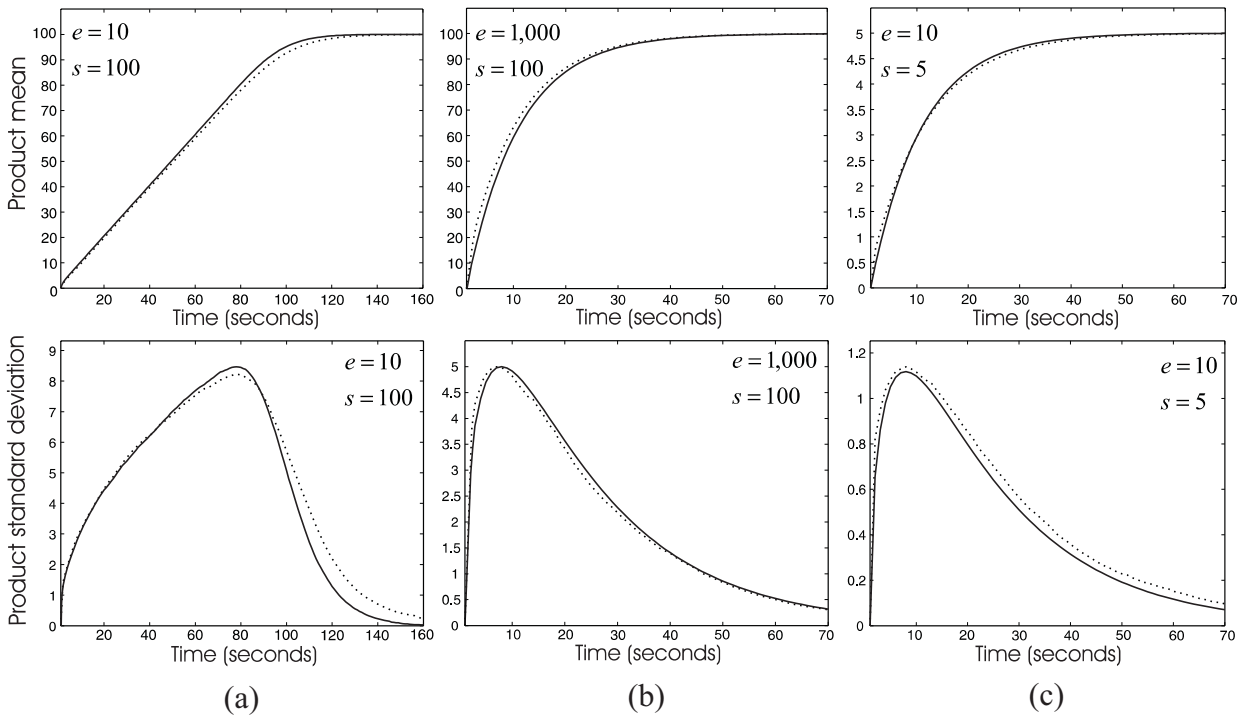

FIG. 1. Exact (dotted lines) and approximate (solid lines) time evolutions of the means and standard deviations of the product process $\{X_1(t), t \geq 0\}$ associated with the "enzyme kinetics" example. We set $c_1 = 0.1\ s^{-1}$, $c_2 = c_3 = 1\ s^{-1}$, and assume the existence of 10 initial enzyme and 100 initial substrate molecules in (a), 1000 initial enzyme and 100 initial substrate molecules in (b), and 10 initial enzyme and 5 initial substrate molecules in (c).

$$\alpha_1(z_1,z_2,z_3) = c_1(-z_1 + z_2 - z_3),$$

$$\alpha_2(z_1,z_2,z_3) = c_2(e + z_1 - z_2 + z_3)(s - z_2 + z_3), \tag{25}$$

$$\alpha_3(z_1,z_2,z_3) = c_3(-z_1 + z_2 - z_3).$$

We assume that the two reactions responsible for the formation of the enzyme/substrate complex (reactions 2 and 3) are fast, as compared to the first reaction, which is slow. In this case, $\mathbf{z}_s = z_1$ and $\mathbf{z}_f = [z_2\ z_3]'$. From Eqs. (15) and (25), we have that

$$\alpha_1^{(t)}(z_1) = -c_1 z_1 + c_1[\mu_2(z_1;t) - \mu_3(z_1;t)]. \tag{26}$$

Therefore, to calculate the propensity function $\alpha_1^{(t)}(z_1)$, we need to determine the difference $\mu_2(z_1;t) - \mu_3(z_1;t)$. We can employ Eqs. (16) and (24), together with the non-negativity constraint $E[\mathbf{X}(t) | \mathbf{Z}_s(t) = \mathbf{z}_s] \geq \mathbf{0}$, to show that $\mu_3(z_1;t) + z_1 \leq \mu_2(z_1;t) \leq \mu_3(z_1;t) + \min\{s, e + z_1\}$. Since z_3 is the DA of a fast reaction, whereas z_1 is the DA of the slow reaction, we expect that, at a sufficiently large time t, $\mu_3(z_1;t) \geq \max\{z_1, \min\{s, e + z_1\}\}$; in this case, we may approximately set $\mu_2(z_1;t) - \mu_3(z_1;t) = \min\{s, e + z_1\}$. This result, together with Eq. (26), leads to

$$\alpha_1^{(t)}(z_1) = c_1[\min\{s, e + z_1\} - z_1],$$

which, together with Eq. (9), leads to the following CME for the slow reaction subsystem:

$$\frac{\partial P(z_1;t)}{\partial t} = c_1[\min\{s, e + z_1 - 1\} - z_1 + 1]P(z_1 - 1;t) - c_1[\min\{s, e + z_1\} - z_1]P(z_1;t). \tag{27}$$

When $e \geq s$, $\min\{s, e + z_1\} = s$ and Eq. (27) becomes

$$\frac{\partial P(z_1;t)}{\partial t} = c_1(s - z_1 + 1)P(z_1 - 1;t) - (s - z_1)P(z_1;t). \tag{28}$$

It can be directly verified that the solution of this CME is given by

$$P(z_1;t) = \binom{s}{z_1}(1 - e^{-c_1 t})^{z_1} e^{-c_1(s - z_1)t}, \quad z_1 = 0, 1, \ldots, s, \tag{29}$$

which is a *binomial* distribution. This observation, together with Eq. (24), leads to

$$E[X_1(t)] = s(1 - e^{-c_1 t}),$$

$$\mathrm{Var}[X_1(t)] = s(1 - e^{-c_1 t})e^{-c_1 t}, \tag{30}$$

for $e \geq s$. Unfortunately, when $e < s$, no analytical solution can be obtained for the distribution of $Z_1(t)$, and thus for the mean and variance of $X_1(t)$. In this case, the mean and variance can be estimated by Monte Carlo simulation from samples of $Z_1(t)$ obtained from the CME (27) using the Gillespie algorithm.

The simulations depicted in Figs. 1 and 2 demonstrate the accuracy of the previous approximation. In Fig. 1, we compare the exact time evolutions of the mean and standard deviation of the product process $\{X_1(t), t \geq 0\}$ (dotted lines) to the approximate time evolutions (solid lines). Since $X_1(t) = Z_1(t)$, these evolutions coincide with the ones associated with the DA process $\{Z_1(t), t \geq 0\}$ of the first reaction. In Fig. 2, we compare the exact and approximate PMFs of $Z_1(t)$ at some time t. The exact quantities are estimated by Monte

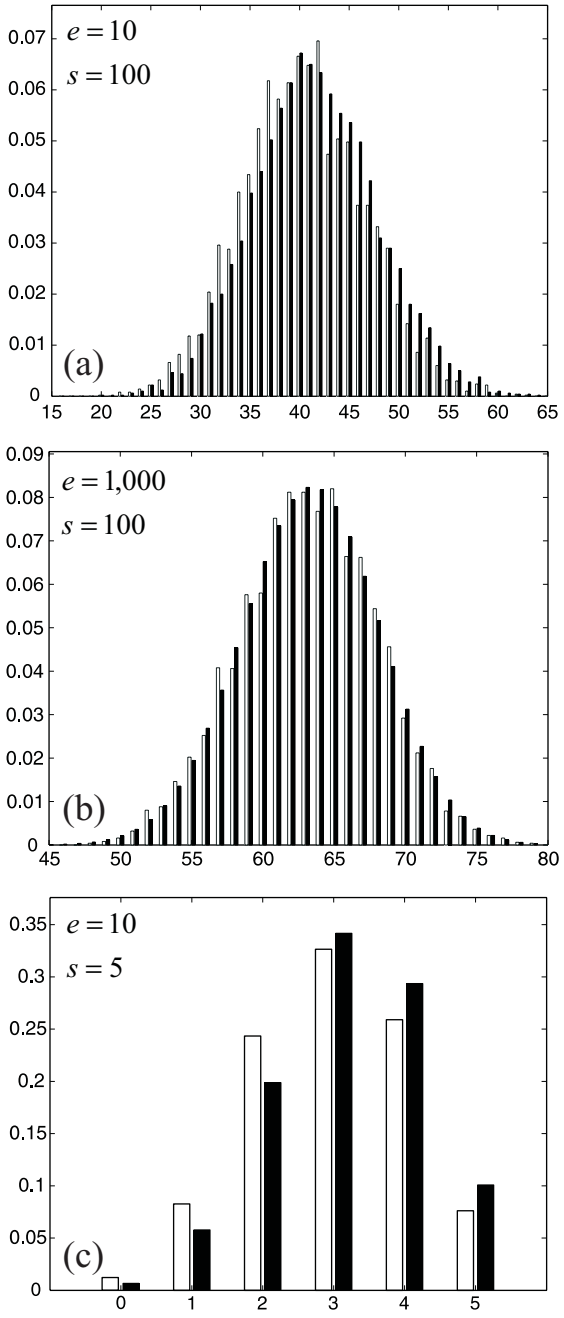

FIG. 2. Exact (white lines) and approximate (black lines) PMFs of the DA process $\{Z_1(t), t \geqslant 0\}$, associated with the "enzyme kinetics" example, at times (a) $t=40$ s, (b) $t=10$ s, and (c) $t=10$ s.

Carlo simulation based on 5000 realizations of $\{Z_1(t), t \geqslant 0\}$, obtained by the Gillespie algorithm applied on the CME (5), with propensity functions given by Eq. (25). When $e \geqslant s$, the approximate PMF and the time evolutions of the mean and standard deviations are determined by Eqs. (29) and (30), whereas when $e < s$, the approximate PMF and time evolutions of the mean and standard deviations are determined by Monte Carlo simulation based on 5000 realizations of $\{Z_1(t), t \geqslant 0\}$, obtained by the Gillespie algorithm applied on the CME (27). These Monte Carlo calculations are orders of magnitude faster than the ones associated with the exact method. The results depicted in Figs. 1 and 2 demonstrate a very good match between the exact and approximate methods.

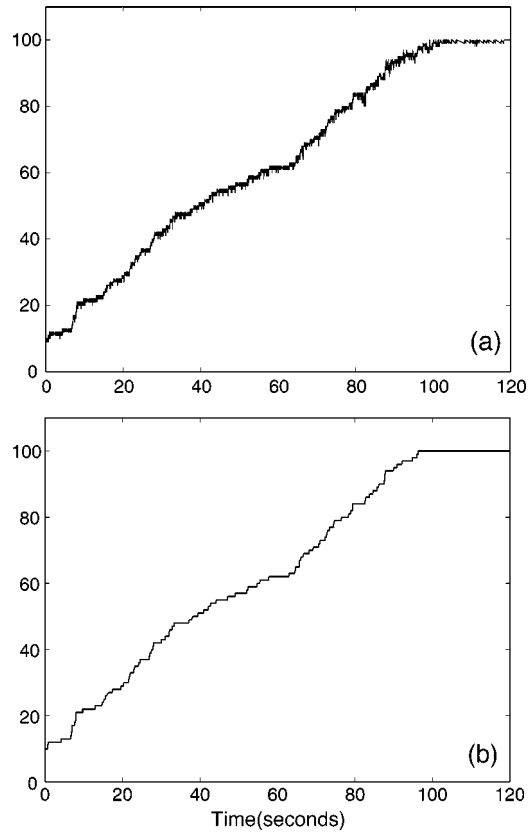

FIG. 3. Typical realizations of (a) the difference $Z_2(t)-Z_3(t)$ and (b) $\min\{s, e+Z_1(t)\}$ in the case of the "enzyme kinetics" example depicted in Fig. 1(a).

To compare our results with the ones presented by Rao and Arkin,[14] we set $c_1=0.1\ \text{s}^{-1}$ and $c_2=c_3=1\ \text{s}^{-1}$. For the results depicted in Fig. 1(a), we assume existence of 10 initial enzyme molecules and 100 initial substrate molecules. These results are in perfect agreement with the ones depicted in Fig. 1 (upper graph) of Rao and Arkin, which are based on a stochastic Michaelis–Menten approximation. For the results depicted in Fig. 1(b), we assume existence of 1000 initial enzyme molecules and 100 initial substrate molecules. These results should be compared with the lower graph of Fig. 1 in Rao and Arkin. Although the stochastic Michaelis–Menten approximation fails to capture the system behavior (since this approximation is based on the assumption that $s \gg e$), our approach provides a very good approximation of the biochemical system by means of the CME (28). Moreover, it provides analytical expressions for the PMF and the mean and variance, given by Eqs. (29) and (30). Simulations indicate that the quality of approximation reduces only slightly in small molecular populations. A typical result for this case is depicted in Fig. 1(c), where we assume existence of ten initial enzyme molecules and only five initial substrate molecules. A comparison of the exact and approximate PMFs yields a very good match as well. Typical comparisons are depicted in Fig. 2 for the three cases considered in Fig. 1. Finally, Fig. 3 depicts typical realizations of $Z_2(t)-Z_3(t)$ and $\min\{s, e+Z_1(t)\}$, for the case when $e=10$ and $s=100$. These results corroborate the validity of replacing z_2-z_3 in $\alpha_1(z_1,z_2,z_3)$ by $\mu_2(z_1;t)-\mu_3(z_1;t)=\min\{s, e+z_1\}$.

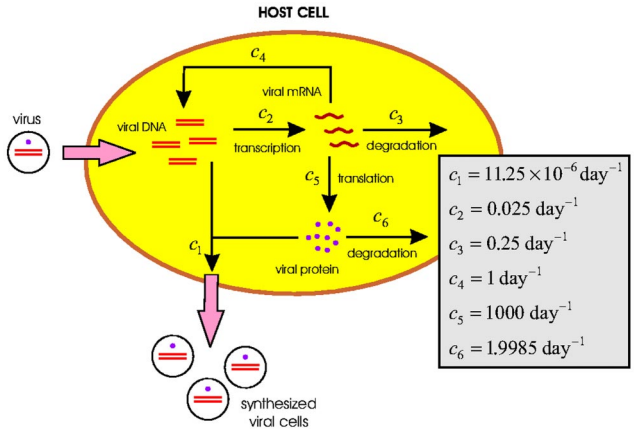

FIG. 4. A simplified model of intracellular viral infection. After a virus is injected into a host cell, the host transcriptional machinery transcribes its DNA into viral mRNA molecules, which are then translated into viral proteins. The viral mRNA is used as a template to replicate the viral DNA. Viral DNA and viral protein molecules are used to assemble multiple copies of the virus, which eventually exit the host cell. Both viral mRNAs and proteins are subject to degradation. The parameters of the model have been obtained from Srivastava and co-workers (Ref. 22).

The method proposed in this paper provides a solution to the classical problem of approximating stochastic enzyme kinetics. Our approach is similar to the one suggested by Rao and Arkin,[14] with three main advantages: (a) when the initial substrate population s is larger than the enzyme population e, we do not need to assume that $s \gg e$ in order to derive the approximation governed by the CME (27), which is the main assumption underlying the derivation of the stochastic Michaelis–Menten approximation in Rao and Arkin; (b) when $e \geqslant s$, the approximation governed by the CME (28) provides an analytical expression for the probability distribution of the DA $Z_1(t)$; and (c) the approximation governed by the CME (27) is accurate, regardless of the values of e and s.

V. EXAMPLE: VIRAL INFECTION KINETICS

To compare our approach to the one proposed by Haseltine and Rawlings,[16] we consider the intracellular viral infection model proposed by Srivastava and co-workers[22] and illustrated in Fig. 4. Based on a set of assumptions and simplifications, the system is governed by the following six reactions,[16,22]

$$\begin{aligned}
&\text{reaction 1:} \quad \mathrm{DNA} + \mathrm{P} \xrightarrow{c_1} \mathrm{V},\\
&\text{reaction 2:} \quad \mathrm{DNA} \xrightarrow{c_2} \mathrm{RNA} + \mathrm{DNA},\\
&\text{reaction 3:} \quad \mathrm{RNA} \xrightarrow{c_3} \varnothing,\\
&\text{reaction 4:} \quad \mathrm{RNA} \xrightarrow{c_4} \mathrm{DNA} + \mathrm{RNA},\\
&\text{reaction 5:} \quad \mathrm{RNA} \xrightarrow{c_5} \mathrm{P} + \mathrm{RNA},\\
&\text{reaction 6:} \quad \mathrm{P} \xrightarrow{c_6} \varnothing.
\end{aligned} \tag{31}$$

We denote by DNA, RNA, and P the viral DNA, mRNA, and protein, respectively. Moreover, we denote by V the number of synthesized viral cells. The first reaction produces viral cells from viral DNA and protein. Although it is most often the case that multiple copies of a protein, or a few different proteins, are used to compose a viral cell,[23] for simplicity, we assume here that a viral cell is composed of one viral DNA and one viral protein molecule. The second and fifth reactions in Eq. (31) model transcription and translation, respectively, whereas the fourth reaction models replication of a viral mRNA template into a viral DNA. These are enzyme-catalyzed reactions, catalyzed by viral DNAs (second reaction), viral mRNAs (fourth and fifth reactions), and associated "cofactors" (e.g., general transcription factors and ribosomes), in which nucleotides and amino acids are the substrates. If we assume that the substrates are available at constant populations, we can approximate transcription, translation, and DNA replication by the monomolecular reactions 2, 4, and 5 in Eq. (31). The third and sixth reactions model viral mRNA and protein degradation, respectively. Finally, the values of the specific probability rate constants are given in Fig. 4.[24]

We use variables $X_n(t)$, $n=1,2,3,4$, to characterize the molecular state of the biochemical system at time $t \geqslant 0$, where

$$X_n \leftrightarrow \begin{cases} \mathrm{V} & \text{for } n=1\\ \mathrm{RNA} & \text{for } n=2\\ \mathrm{DNA} & \text{for } n=3\\ \mathrm{P} & \text{for } n=4. \end{cases}$$

From Eqs. (2) and (31), we have

$$\begin{aligned}
h_1(x_1,x_2,x_3,x_4) &= x_3x_4,\\
h_2(x_1,x_2,x_3,x_4) &= x_3,\\
h_3(x_1,x_2,x_3,x_4) &= x_2,\\
h_4(x_1,x_2,x_3,x_4) &= x_2,\\
h_5(x_1,x_2,x_3,x_4) &= x_2,\\
h_6(x_1,x_2,x_3,x_4) &= x_4,
\end{aligned} \tag{32}$$

whereas the stoichiometry matrix is given by

$$\mathbb{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & -1 & 0 & 0 & 0\\ -1 & 0 & 0 & 1 & 0 & 0\\ -1 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}. \tag{33}$$

By following Srivastava and co-workers[22] and Haseltine and Rawlings,[16] we assume that the host cell is initially infected by one viral mRNA molecule; in this case we set

$$\mathbf{x}_0 = [0 \quad 1 \quad 0 \quad 0]'. \tag{34}$$

From Eqs. (4), (33), and (34), we have

$$
\begin{aligned}
X_1(t) &= Z_1(t),\\
X_2(t) &= 1 + Z_2(t) - Z_3(t),\\
X_3(t) &= -Z_1(t) + Z_4(t),\\
X_4(t) &= -Z_1(t) + Z_5(t) - Z_6(t).
\end{aligned}
\tag{35}
$$

Finally, note that the CME underlying the stochastic evolution of the associated DAs is given by Eq. (5), with propensity functions [see Eqs. (6), (32), and (35)]

$$
\begin{aligned}
\alpha_1(\mathbf{z}) &= c_1(-z_1 + z_4)(-z_1 + z_5 - z_6),\\
\alpha_2(\mathbf{z}) &= c_2(-z_1 + z_4),\\
\alpha_3(\mathbf{z}) &= c_3(1 + z_2 - z_3),\\
\alpha_4(\mathbf{z}) &= c_4(1 + z_2 - z_3),\\
\alpha_5(\mathbf{z}) &= c_5(1 + z_2 - z_3),\\
\alpha_6(\mathbf{z}) &= c_6(-z_1 + z_5 - z_6).
\end{aligned}
\tag{36}
$$

It has been noted by Haseltine and Rawlings[16] that, when the number of viral mRNAs is greater than zero and the number of viral proteins is more than 100, the last two reactions in Eq. (31) occur much faster than reactions 1–4. Therefore, we set $\mathbf{z}_s=[z_1\ z_2\ z_3\ z_4]'$ and $\mathbf{z}_f=[z_5\ z_6]'$; in this case [recall Eqs. (15) and (36)]

$$
\begin{aligned}
\alpha_1^{(t)}(\mathbf{z}_s) &= c_1(-z_1 + z_4)[\mu_5(\mathbf{z}_s;t) - \mu_6(\mathbf{z}_s;t) - z_1],\\
\alpha_2^{(t)}(\mathbf{z}_s) &= c_2(-z_1 + z_4),\\
\alpha_3^{(t)}(\mathbf{z}_s) &= c_3(1 + z_2 - z_3),\\
\alpha_4^{(t)}(\mathbf{z}_s) &= c_4(1 + z_2 - z_3).
\end{aligned}
\tag{37}
$$

Note that approximation of the fast reactions 5 and 6 affects only the propensity function of viral cell synthesis (first reaction). Computation of the new propensity function requires calculation of the difference $\mu_5(\mathbf{z}_s;t)-\mu_6(\mathbf{z}_s;t)$. In contrast to the previous example, we cannot derive an approximation to $\mu_5(\mathbf{z}_s;t)-\mu_6(\mathbf{z}_s;t)$ by only using Eq. (35) and the nonnegativity constraint $E[\mathbf{X}(t)\,|\,\mathbf{Z}_f(t)=\mathbf{z}_f]\geqslant\mathbf{0}$.

Since, within the coarse time scale, the fast reaction subsystem is highly reactive, we may assume that, given the state of the slow reaction subsystem, the fast reaction subsystem rapidly reaches a state of equilibrium. This implies that, for all practical purposes, the probability of a viral protein to be synthesized during a sufficiently small time interval $[t,t+dt)$ by reaction 5 will approximately equal the probability of a viral protein to be degraded during $[t,t+dt)$ by reaction 6. This sets the well-known quasiequilibrium assumption of deterministic chemical kinetics[15] in a statistical framework. For this reason, we refer to this assumption as the stochastic quasiequilibrium assumption (SQEA).

Recall now that the probability of a viral protein to be synthesized during a sufficiently small time interval $[t,t+dt)$ is given by $a_5(\mathbf{Z}_s(t),\mathbf{Z}_f(t))dt$, whereas the probability of a viral protein to be degraded during that interval is given by $a_6(\mathbf{Z}_s(t),\mathbf{Z}_f(t))dt$. Therefore, the SQEA leads to $a_5(\mathbf{Z}_s(t),\mathbf{Z}_f(t))\simeq a_6(\mathbf{Z}_s(t),\mathbf{Z}_f(t))$, which, together with Eq. (36), results in $Z_5(t)-Z_6(t)\simeq Z_1(t)+(c_5/c_6)[1+Z_2(t)-Z_3(t)]$. By taking conditional expectations on both sides of this equation, we approximately obtain $\mu_5(\mathbf{z}_s;t)-\mu_6(\mathbf{z}_s;t)=z_1+(c_5/c_6)(1+z_2-z_3)$, which, together with Eq. (37), results in

$$
\alpha_1^{(t)}(\mathbf{z}_s) = \frac{c_1c_5}{c_6}(-z_1 + z_4)(1 + z_2 - z_3). \tag{38}
$$

Figure 5 depicts the estimated PMF of viral mRNA molecules at time $t=200$ days, whereas Fig. 6 depicts the dynamic evolutions of the mean (solid curves) and standard deviations (dotted curves) of the four processes $\{X_n(t),t\geqslant 0\}$, $n=1,2,3,4$, where $X_1(t)$, $X_2(t)$, $X_3(t)$ are given byEq. (35), and $X_4(t)=\lfloor(c_5/c_6)[1+Z_2(t)-Z_3(t)]\rfloor$, with $\lfloor\ \rfloor$ being the *floor* function. These results have been estimated by Monte Carlo simulation based on 4000 DA realizations, obtained by the Gillespie algorithm applied on the CME (9), with propensity functions given by Eqs. (37) and (38). The results depicted in Figs. 5 and 6 compare most favorably with results reported by Haseltine and Rawlings (see Figs. 7–10),[16] obtained by exact and approximate Monte Carlo simulations, and demonstrate the accuracy of the proposed technique for approximating fast reaction kinetics in the intracellular viral infection example.

VI. EXAMPLE: TRANSCRIPTION REGULATION

In a final example, we consider the simple transcriptional regulatory system depicted in Fig. 7. The following ten reactions characterize the system:

$$
\begin{aligned}
&\text{reaction 1:}\quad \mathrm{RNA}\xrightarrow{c_1}\mathrm{RNA}+\mathrm{M},\\
&\text{reaction 2:}\quad \mathrm{M}\xrightarrow{c_2}\varnothing,\\
&\text{reaction 3:}\quad \mathrm{DNA\cdot D}\xrightarrow{c_3}\mathrm{RNA}+\mathrm{DNA\cdot D},\\
&\text{reaction 4:}\quad \mathrm{RNA}\xrightarrow{c_4}\varnothing,\\
&\text{reaction 5:}\quad \mathrm{DNA}+\mathrm{D}\xrightarrow{c_5}\mathrm{DNA\cdot D},\\
&\text{reaction 6:}\quad \mathrm{DNA\cdot D}\xrightarrow{c_6}\mathrm{DNA}+\mathrm{D},\\
&\text{reaction 7:}\quad \mathrm{DNA\cdot D}+\mathrm{D}\xrightarrow{c_7}\mathrm{DNA\cdot 2D},\\
&\text{reaction 8:}\quad \mathrm{DNA\cdot 2D}\xrightarrow{c_8}\mathrm{DNA\cdot D}+\mathrm{D},\\
&\text{reaction 9:}\quad \mathrm{M}+\mathrm{M}\xrightarrow{c_9}\mathrm{D},\\
&\text{reaction 10:}\quad \mathrm{D}\xrightarrow{c_{10}}\mathrm{M}+\mathrm{M},
\end{aligned}
\tag{39}
$$

where M is the protein (monomer), D is the transcription factor (dimer), DNA is the DNA template free of dimers, DNA·D is the DNA template bound at $\mathcal{R}_1$, DNA·2D is the DNA template bound at $\mathcal{R}_1$ and $\mathcal{R}_2$, and RNA is the mRNA

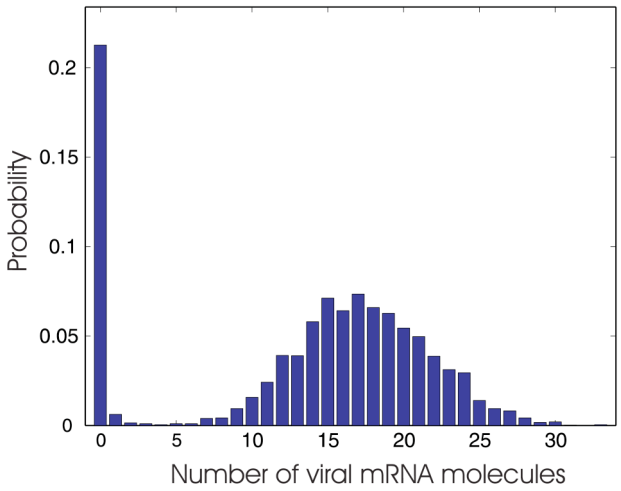

FIG. 5. Estimated PMF of viral mRNA molecules, associated with the "intracellular viral infection" example, at time $t=200$ days.

produced by transcription. The first reaction in Eq. (39) models translation of mRNA into protein M, whereas reaction 3 models transcription. Moreover, reactions 2 and 4 model degradation of M and *m*RNA, respectively. We assume that transcription occurs only when D occupies the binding site $\mathcal{R}_1$ and that the gene is not subject to basal transcription. Reactions 5–8 model dimer/DNA binding/unbinding. We assume that D may occupy the binding site $\mathcal{R}_2$ only when $\mathcal{R}_1$ is occupied by D (cooperative binding). Finally, reactions 9 and 10 model dimerization of M to D.

We use variables $X_n(t)$, $n=1,2,\ldots,6$, to characterize the molecular state of the transcriptional regulatory system at time $t \geqslant 0$, where

$$X_n \leftrightarrow \begin{cases} \text{M} & \text{for } n=1 \\ \text{D} & \text{for } n=2 \\ \text{RNA} & \text{for } n=3 \\ \text{DNA} & \text{for } n=4 \\ \text{DNA} \cdot \text{D} & \text{for } n=5 \\ \text{DNA} \cdot \text{2D} & \text{for } n=6. \end{cases}$$

From Eqs. (2) and (39), we have

$$\begin{aligned} h_1(x_1,x_2,x_3,x_4,x_5,x_6) &= x_3, \\ h_2(x_1,x_2,x_3,x_4,x_5,x_6) &= x_1, \\ h_3(x_1,x_2,x_3,x_4,x_5,x_6) &= x_5, \\ h_4(x_1,x_2,x_3,x_4,x_5,x_6) &= x_3, \\ h_5(x_1,x_2,x_3,x_4,x_5,x_6) &= x_2x_4, \\ h_6(x_1,x_2,x_3,x_4,x_5,x_6) &= x_5, \\ h_7(x_1,x_2,x_3,x_4,x_5,x_6) &= x_2x_5, \\ h_8(x_1,x_2,x_3,x_4,x_5,x_6) &= x_6, \end{aligned} \tag{40}$$

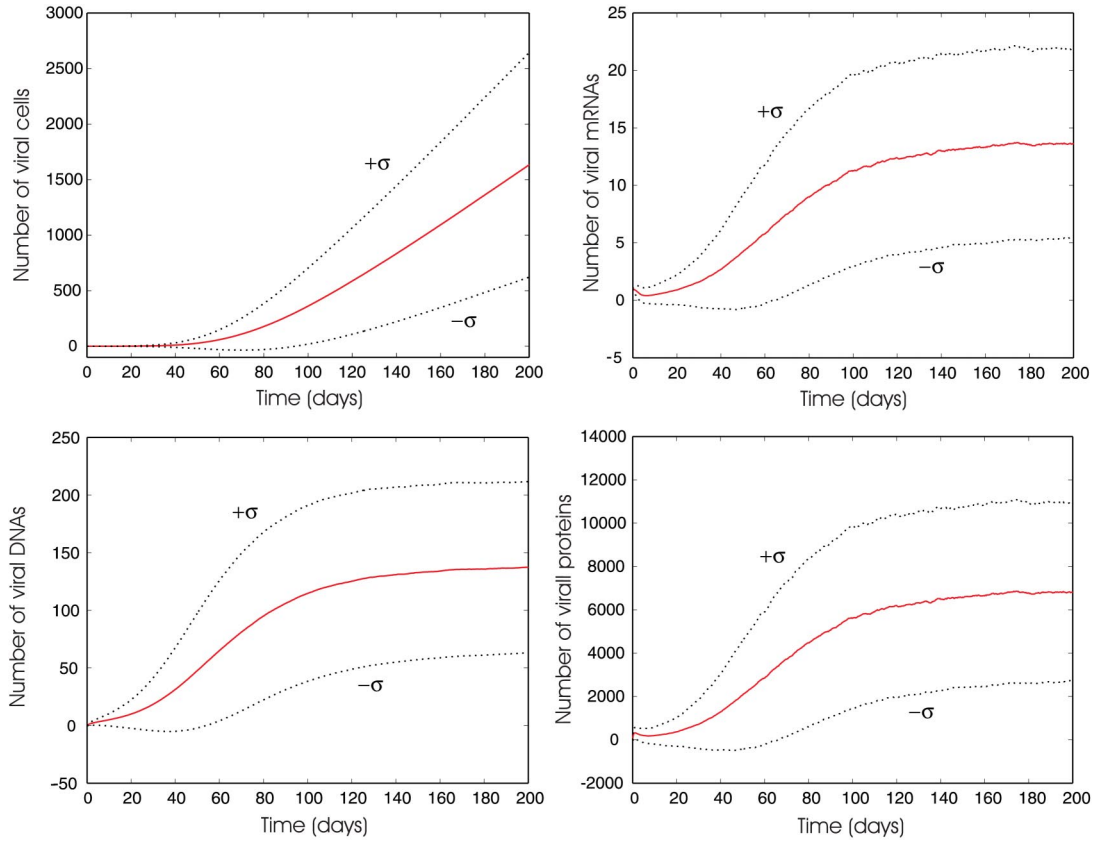

FIG. 6. Estimated dynamic evolutions of the mean (solid lines) and standard deviations (dotted lines) of the viral cell, mRNA, DNA, and protein populations, associated with the "intracellular viral infection" example, during a period of 200 days.

$$h_9(x_1,x_2,x_3,x_4,x_5,x_6) = x_1(x_1-1)/2,$$

$$h_{10}(x_1,x_2,x_3,x_4,x_5,x_6) = x_2,$$

whereas the stoichiometry matrix is given by

$$\mathbb{S} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 2 \\ 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}. \tag{41}$$

We initialize the system with $m \geqslant 1$ monomers and $d \geqslant 1$ dimers, and assume $g \geqslant 1$ DNA templates per cell; in this case, we have

$$\mathbf{x}_0 = [m \quad d \quad 0 \quad g \quad 0 \quad 0]'. \tag{42}$$

From Eqs. (4), (41), and (42), note that

$$\begin{aligned} X_1(t) &= m + Z_1(t) - Z_2(t) - 2Z_9(t) + 2Z_{10}(t), \\ X_2(t) &= d - Z_5(t) + Z_6(t) - Z_7(t) + Z_8(t) + Z_9(t) - Z_{10}(t), \\ X_3(t) &= Z_3(t) - Z_4(t), \\ X_4(t) &= g - Z_5(t) + Z_6(t), \\ X_5(t) &= Z_5(t) - Z_6(t) - Z_7(t) + Z_8(t), \\ X_6(t) &= Z_7(t) - Z_8(t). \end{aligned} \tag{43}$$

Finally, the CME underlying the stochastic evolution of the associated DAs is given by Eq. (5) with propensity functions [see Eqs. (6), (40), and (43)]

$$\begin{aligned} \alpha_1(\mathbf{z}) &= c_1(z_3 - z_4), \\ \alpha_2(\mathbf{z}) &= c_2(m + z_1 - z_2 - 2z_9 + 2z_{10}), \\ \alpha_3(\mathbf{z}) &= c_3(z_5 - z_6 - z_7 + z_8), \\ \alpha_4(\mathbf{z}) &= c_4(z_3 - z_4), \\ \alpha_5(\mathbf{z}) &= c_5(d - z_5 + z_6 - z_7 + z_8 + z_9 - z_{10})(g - z_5 + z_6), \\ \alpha_6(\mathbf{z}) &= c_6(z_5 - z_6 - z_7 + z_8), \\ \alpha_7(\mathbf{z}) &= c_7(d - z_5 + z_6 - z_7 + z_8 + z_9 - z_{10}) \\ &\quad \times (z_5 - z_6 - z_7 + z_8), \\ \alpha_8(\mathbf{z}) &= c_8(z_7 - z_8), \\ \alpha_9(\mathbf{z}) &= c_9(m + z_1 - z_2 - 2z_9 + 2z_{10}) \\ &\quad \times (m + z_1 - z_2 - 2z_9 + 2z_{10} - 1)/2, \\ \alpha_{10}(\mathbf{z}) &= c_{10}(d - z_5 + z_6 - z_7 + z_8 + z_9 - z_{10}). \end{aligned} \tag{44}$$

To use biologically relevant parameter values, we employ the system depicted in Fig. 7 to mathematically model transcription regulation of the bacteriophage λ repressor protein (CI), responsible for maintaining lysogeny of the λ virus in *E. coli* cells.[25–29] For simplicity, we consider a mutant version of the system for which the operator site O_R1 is absent from the regulatory region of that gene.[28] In this case, dimers D can only bind at operator sites O_R2 or O_R3, designated as $\mathcal{R}_1$ and $\mathcal{R}_2$ in Fig. 7 (specific binding). Note that our purpose is not to provide a detailed mathematical model for lysogeny in *E. coli* cells but to illustrate the proposed approximation technique with a biologically relevant example.

The values of the specific probability rate constants c_m are listed in Fig. 7. These values are calculated by Eq. (3) from published values of the reaction rate constants k_m. The values of k_1, k_4, and k_5 are obtained from Bundschun and co-workers,[30] whereas the value of k_3 is taken from Santillán and Mackey.[31] Moreover, the values of k_2 and k_{10} are obtained from Table 3 of Ref. 27. We set $k_6 = k_5 \exp\{\Delta G_1/RT\}$, $k_7 = k_5/100$, and $k_8 = k_7 \exp\{\Delta G_2/RT\}$, where $R = 1.9872$ cal mol^{-1} K^{-1} is the gas constant and $T = 310.15$ K is the absolute temperature (37 °C). Moreover, $\Delta G_1 = -10.5$ kcal/mol is the binding energy of D to $\mathcal{R}_1$ and $\Delta G_2 = -22.9$ kcal/mol is the binding energy of D to $\mathcal{R}_2$.[32] We assume that the volume of an *E. coli* cell grows between successive cell divisions in accordance to $V(t) = V(0)e^{\ln(2)t/\tau}$, where $V(0)$ is the initial cellular volume and τ is the average cell cycle time.[33] We set $V(0) = 10^{-15}$ l and $\tau = 35$ min.[27] For simplicity, we do not model cell division and limit our simulations to within the 35 min cell cycle time period. We finally set $m = 2$, $d = 6$, and assume that an average *E. coli* cell is infected by about two λ chromosomes (i.e., $g = 2$).[34]

It is well known that dimerization is a fast reaction, as compared to transcription, translation, mRNA and protein degradation, and protein/DNA binding/unbinding. Therefore, we set $\mathbf{z}_s = [z_1\, z_2\, z_3\, z_4\, z_5\, z_6\, z_7\, z_8]'$ and $\mathbf{z}_f = [z_9\, z_{10}]'$; in this case [recall Eqs. (15) and (44)]

$$\begin{aligned} \alpha_1^{(t)}(\mathbf{z}) &= c_1(z_3 - z_4), \\ \alpha_2^{(t)}(\mathbf{z}) &= c_2[m + z_1 - z_2 - 2\mu_9(\mathbf{z}_s;t) + 2\mu_{10}(\mathbf{z}_s;t)], \\ \alpha_3^{(t)}(\mathbf{z}) &= c_3(z_5 - z_6 - z_7 + z_8), \\ \alpha_4^{(t)}(\mathbf{z}) &= c_4(z_3 - z_4), \\ \alpha_5^{(t)}(\mathbf{z}) &= c_5[d - z_5 + z_6 - z_7 + z_8 + \mu_9(\mathbf{z}_s;t) - \mu_{10}(\mathbf{z}_s;t)] \\ &\quad \times (g - z_5 + z_6), \\ \alpha_6^{(t)}(\mathbf{z}) &= c_6(z_5 - z_6 - z_7 + z_8), \\ \alpha_7^{(t)}(\mathbf{z}) &= c_7[d - z_5 + z_6 - z_7 + z_8 + \mu_9(\mathbf{z}_s;t) - \mu_{10}(\mathbf{z}_s;t)] \\ &\quad \times (z_5 - z_6 - z_7 + z_8), \\ \alpha_8^{(t)}(\mathbf{z}) &= c_8(z_7 - z_8). \end{aligned} \tag{45}$$

Note that approximation of the fast reactions 9 and 10 affects only the propensity functions of monomer degradation (sec-

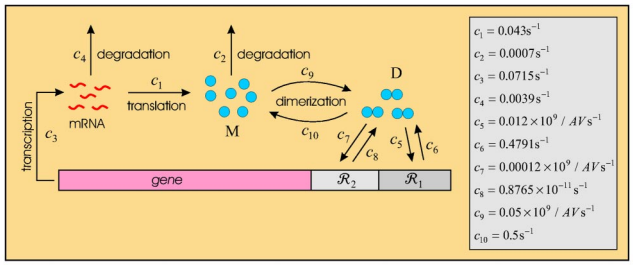

FIG. 7. A simple transcriptional regulatory system. Protein M, synthesized by transcription of a gene, dimerizes to the transcription factor D, which may bind to the gene's regulatory region at two binding sites, $\mathcal{R}_1$ and $\mathcal{R}_2$. Binding of D at $\mathcal{R}_1$ activates transcription of M. However, binding of D at $\mathcal{R}_2$ excludes the RNA polymerase from binding at the gene's promoter; in this case transcription is repressed.

ond reaction) and dimer/DNA binding (fifth and seventh reactions). Computation of the new propensity functions requires calculation of the difference $\mu_9(\mathbf{z}_s;t)-\mu_{10}(\mathbf{z}_s;t)$, which can be computed by using the SQEA employed in the previous example. Indeed, the probability that a dimer D is synthesized during a sufficiently small time interval $[t,t+dt)$ is given by $\alpha_9(\mathbf{Z}_s(t),\mathbf{Z}_f(t))dt$, whereas the probability that a dimer D is decomposed into two monomers M during the same time interval is given by $\alpha_{10}(\mathbf{Z}_s(t),\mathbf{Z}_f(t))dt$. In this case, the SQEA leads to $\alpha_9(\mathbf{Z}_s(t),\mathbf{Z}_f(t))\simeq\alpha_{10}(\mathbf{Z}_s(t),\mathbf{Z}_f(t))$, which, together with Eq. (44), approximately results in

$$[Z_9(t)-Z_{10}(t)]^2-A(\mathbf{Z}_s(t))[Z_9(t)-Z_{10}(t)]+B(\mathbf{Z}_s(t))=0,$$

where

$$A(\mathbf{z}_s)=m+z_1-z_2+\frac{c_{10}}{2c_9}-\frac{1}{2},$$

$$B(\mathbf{z}_s)=\tfrac{1}{4}(m+z_1-z_2)(m+z_1-z_2-1)-\frac{c_{10}}{2c_9}(d-z_5+z_6-z_7+z_8).$$

By solving the previous quadratic equation and by taking conditional expectations with respect to $\mathbf{Z}_s(t)=\mathbf{z}_s$, we obtain

$$\mu_9(\mathbf{z}_s;t)-\mu_{10}(\mathbf{z}_s;t)=\tfrac{1}{2}[A(\mathbf{z}_s)\pm\sqrt{A^2(\mathbf{z}_s)-4B(\mathbf{z}_s)}]. \quad (46)$$

Moreover, from Eq. (43), the non-negativity constraints $E[X_1(t)|\mathbf{Z}_s(t)=\mathbf{z}_s]\geqslant 0$ and $E[X_2(t)|\mathbf{Z}_s(t)=\mathbf{z}_s]\geqslant 0$, and some straightforward algebra, we can show that $A^2(\mathbf{z}_s)\geqslant 4B(\mathbf{z}_s)$, which implies that the difference $\mu_9(\mathbf{z}_s;t)-\mu_{10}(\mathbf{z}_s;t)$, given by Eq. (46), is *always* real valued. Note that Eq. (46), together with the non-negativity constraint $E[X_1(t)|\mathbf{Z}_s(t)=\mathbf{z}_s]\geqslant 0$, implies that $\mu_9(\mathbf{z}_s;t)-\mu_{10}(\mathbf{z}_s;t)\leqslant 0.5[m+z_1-z_2]$, from which we obtain

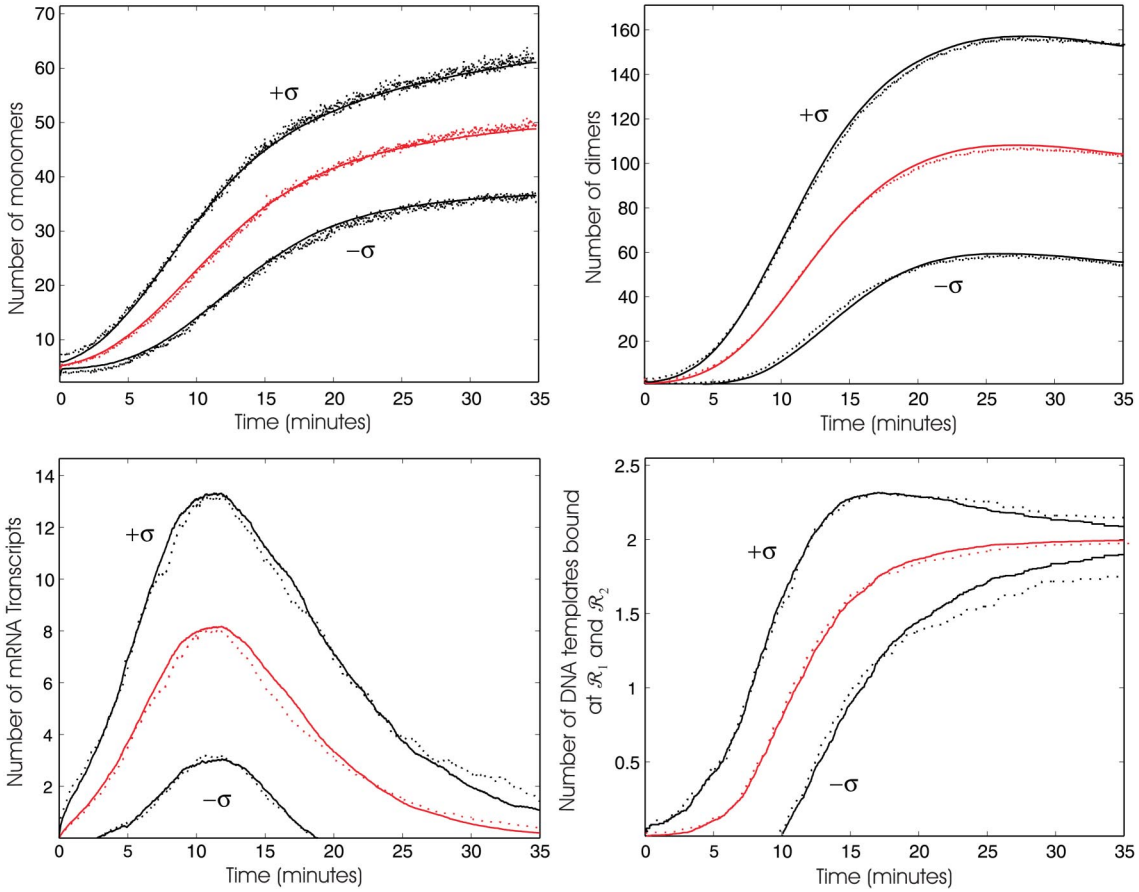

FIG. 8. Exact (dotted lines) and approximate (solid lines) dynamic evolutions of the mean and standard deviations of the number of monomers, dimers, mRNA transcripts, and DNA templates bound at regulatory regions $\mathcal{R}_1$ and $\mathcal{R}_2$, associated with the "transcription regulation" example, during a period of 35 min.

$$\pm\sqrt{A^2(\mathbf{z}_s)-4B(\mathbf{z}_s)} \leqslant \frac{1}{2}\frac{c_9-c_{10}}{c_9}. \tag{47}$$

However, since $c_9 \leqslant c_{10}$, Eq. (47) can be satisfied only when the $-$ sign is used on the left-hand side. Therefore, we set $\mu_9(\mathbf{z}_s;t)-\mu_{10}(\mathbf{z}_s;t)=0.5[A(\mathbf{z}_s)-\sqrt{A^2(\mathbf{z}_s)-4B(\mathbf{z}_s)}]$, which, together with Eq. (45), results in

$$\begin{aligned}
\alpha_2^{(t)}(\mathbf{z}) &= c_2[m+z_1-z_2-A(\mathbf{z}_s)+\sqrt{A^2(\mathbf{z}_s)-4B(\mathbf{z}_s)}],\\
\alpha_5^{(t)}(\mathbf{z}) &= c_5\Big\{d-z_5+z_6-z_7+z_8+\tfrac{1}{2}[A(\mathbf{z}_s)\\
&\quad -\sqrt{A^2(\mathbf{z}_s)-4B(\mathbf{z}_s)}]\Big\}(g-z_5+z_6),\\
\alpha_7^{(t)}(\mathbf{z}) &= c_7\Big[d-z_5+z_6-z_7+z_8+\tfrac{1}{2}\{A(\mathbf{z}_s)\\
&\quad -\sqrt{A^2(\mathbf{z}_s)-4B(\mathbf{z}_s)}\}\Big](z_5-z_6-z_7+z_8).
\end{aligned} \tag{48}$$

An important observation is that approximation of the two fast dimerization reactions requires knowledge of the ratio c_9/c_{10} of the two specific probability rate constants c_9 and c_{10}. This ratio is much easier to determine experimentally than the actual values of c_9 and c_{10}. Therefore, in addition to reducing computational complexity, the proposed technique reduces model complexity in a fashion similar to that discussed by Rao and Arkin.[14]

Figure 8 depicts the exact (dotted lines) and approximate (solid lines) dynamic evolutions of the means and standard deviations of the numbers of monomers, dimers, mRNA transcripts, and DNA templates bound at regulatory regions $\mathcal{R}_1$ and $\mathcal{R}_2$. The exact results have been estimated with Monte Carlo sampling based on 200 DA realizations, obtained by the Gillespie algorithm applied on the CME (5), with propensity functions given by Eq. (44). The approximate results have been estimated with Monte Carlo sampling based on 1000 DA realizations, obtained by the Gillespie algorithm, modified for correctly simulating the chemical reactions due to growth of the cellular volume,[35] applied on the CME (8), with propensity functions given by Eqs. (45) and (48). Once more, these results demonstrate the accuracy of the proposed technique for approximating fast reaction kinetics. Gillespie simulation of the approximate CME is orders of magnitude faster than Gillespie simulation of the exact CME. Typically, a 35 min simulation, coded in Matlab®, takes about 1–3 h on a dual-processor PC running Windows 2000 for the exact case, and a mere 1 s for the approximate case.

VII. DISCUSSION

The technique presented in this paper is based on our ability to classify the reactions of a stochastic biochemical system into slow and fast reactions. Practical issues addressing this problem have been discussed by Haseltine and Rawlings.[16] By using the degrees of advancement to characterize the state of a stochastic biochemical system, we have shown that we can effectively partition the system into a slow and a fast reaction subsystem. The slow reaction subsystem is characterized by the CME (9), which requires knowledge of the propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, given by Eq. (10). When the slow reaction subsystem consists of monomolecular reactions, bimolecular reactions with identical reactants that are not influenced by fast reactions, or bimolecular reactions with different reactants for which at most one reactant is influenced by fast reactions, the propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, are given by Eq. (15). In this case, the problem of eliminating fast reaction kinetics boils down to the problem of calculating the conditional mean DAs of the fast reactions, given the state of the slow reaction subsystem. The examples presented in Secs. IV–VI, show that the slow reactions of several biological mechanisms satisfy the previous requirements.

By ignoring the third term on the right-hand side of Eq. (17), we may be able to use Eq. (15) in cases when both reactants of a slow bimolecular reaction are influenced by fast reactions. However, this approximation may introduce substantial errors, depending on the magnitudes of the second-order partial derivatives of the propensity functions $a_m(\mathbf{z}_s,\mathbf{z}_f)$, $m \in \mathcal{M}_s$, with respect to $\mathbf{z}_f$ and on the value of the conditional covariance matrix of the fast DA vector $\mathbf{Z}_f$, given the state of the slow reaction subsystem.

A more desirable approach will be to compute the propensity functions $a_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, from Eqs. (10) and (17), by calculating the conditional mean of the third term on the right-hand side of Eq. (17). This problem is, however, very difficult, and certain approximations are needed in order to proceed. One possibility is to set $\mathbb{S}=[\mathbb{S}_s\ \mathbb{S}_f]$, which provides a partition of the stoichiometry matrix into slow and fast reaction submatrices, $\mathbb{S}_s$ and $\mathbb{S}_f$, respectively, so that $\mathbf{X}(t)=\mathbf{x}_0+\mathbb{S}_s\mathbf{Z}_s(t)+\mathbb{S}_f\mathbf{Z}_f(t)$ [recall Eq. (4)]. In this case

$$\begin{aligned}
&(\mathbf{Z}_f-\mu)'\frac{\partial^2\alpha_m(\mathbf{z}_s,\mu)}{\partial\mathbf{z}_f^2}(\mathbf{Z}_f-\mu)\\
&\quad=(\mathbf{X}_f-\mathbb{S}_f\mu)'\frac{\partial^2\alpha_m(\mathbf{z}_s,\mu)}{\partial\mathbf{x}_f^2}(\mathbf{X}_f-\mathbb{S}_f\mu),
\end{aligned} \tag{49}$$

where $\mathbf{X}_f \triangleq \mathbb{S}_f\mathbf{Z}_f$. Note that Eqs. (2) and (6) imply that the second-order partial derivative of the propensity function $a_m(\mathbf{z}_s,\mathbf{z}_f)$ of a bimolecular reaction with respect to $\mathbf{x}_f$ is either zero, c_m, or $2c_m$, which leaves us with the task of calculating the conditional variances and covariances of $\mathbf{X}_f$. We may obtain a reasonable *approximation* by assuming that the conditional covariances of $\mathbf{X}_f$ are negligible and that, given the state of the slow reaction subsystem, the fast reactions are mutually independent, with the number of occurrences of the m^{th} fast reaction at time t following a Poisson distribution with parameter $\mu_m(\mathbf{z}_s;t)$, given by Eq. (16). In this case, since $\mathbf{X}_f$ is a linear combination of mutually independent Poisson random variables in $\mathbf{Z}_f$, each element of $\mathbf{X}_f$ will be a Poisson random variable as well with mean and variance given by the corresponding element in $\mathbb{S}_f\mu$. It is clear that, with this approach, we only need to know the conditional mean vectors $\mathbb{S}_f\mu_m(\mathbf{z}_s;t)$, $m \in \mathcal{M}_f$, in order to obtain an approximation of the propensity functions $a_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$.

The examples presented in this paper illustrate an effective technique for estimating the conditional mean vectors $\mathbb{S}_f\mu_m(\mathbf{z}_s;t)$, $m \in \mathcal{M}_f$, which employs the SQEA discussed in Sec. V in conjunction with the non-negativity constraint $E[\mathbf{X}(t)\,|\,\mathbf{Z}_s(t)=\mathbf{z}_s] \geqslant 0$. The simplicity of the fast reaction subsystems considered in the examples allowed us to derive analytical expressions for $\mathbb{S}_f\mu_m(\mathbf{z}_s;t)$, $m \in \mathcal{M}_f$. This may not be

possible in more complex cases. However, we may be able to employ the same technique, together with a computational method, to numerically evaluate these means. The price to be paid will be an increase in computational complexity. In the following, we briefly discuss a possible computational method.

In general, the product of a fast reaction must be reactant in at least one other fast reaction. Otherwise, the number of molecules produced by fast reactions will eventually grow to infinity. For a fast reaction $m \in \mathcal{M}_f$ with product P, let us denote by $\mathcal{N}(m)$ all fast reactions with reactant P. Given that $\mathbf{Z}_s(t)=\mathbf{z}_s$, the application of the SQEA produces the following set of equations:

$$a_m(\mathbf{z}_s,\mu) = \sum_{n\in\mathcal{N}(m)} a_n(\mathbf{z}_s,\mu), \quad m \in \mathcal{M}_f,$$

which, together with Eq. (6), implies that

$$h_m(\mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u}) = \sum_{n\in\mathcal{N}(m)} \frac{c_n}{c_m} h_n(\mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u}), \quad m \in \mathcal{M}_f, \tag{50}$$

where $\mathbf{u} \triangleq \mathbb{S}_f\mu$. Since $a_m^{(t)}(\mathbf{z}_s)=c_m h_m(\mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u})$, $m \in \mathcal{M}_s$, we need to find a vector $\mathbf{u}$ that satisfies Eq. (50), subject to the non-negativity constraint $E[\mathbf{X}(t)\,|\,\mathbf{Z}_s(t)=\mathbf{z}_s]=\mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u} \geqslant 0$. Therefore, the problem of calculating the propensity functions $a_m^{(t)}(\mathbf{z}_s)$ reduces to the problem of finding a solution $\mathbf{u}$ to the following system of equations:

$$\mathbb{C}\mathbf{h}(\mathbf{u}) = \mathbf{0} \quad \text{subject to} \quad \mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u} \geqslant \mathbf{0}, \tag{51}$$

where $\mathbb{C}$ is an $(M-M_0)\times(M-M_0)$ matrix whose diagonal elements are 1 and the remaining elements are either 0 or $-c_n/c_m$, $m,\ n \in \mathcal{M}_f$, and $\mathbf{h}(\mathbf{u})$ is an $(M-M_0)\times 1$ vector whose m^{th} element is $h_m(\mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u})$, $m \in \mathcal{M}_f$.

The system in Eq. (51) comprises at most quadratic equations subject to linear inequality constraints. We can reformulate this problem to obtain an unconstrained system of equations that can be numerically solved by means of the well-known *Newton–Raphson method*.[36] We can do so by employing the *barrier method*.[37] Indeed, $\mathbf{u}$ satisfies Eq. (51) if and only if

$$\mathbb{C}\mathbf{h}(\mathbf{u}) + \sum_{m\in\mathcal{M}} I_+[e_m(\mathbf{u})] = \mathbf{0}, \tag{52}$$

where $\mathbf{e}(\mathbf{u}) \triangleq \mathbf{x}_0+\mathbb{S}_s\mathbf{z}_s+\mathbf{u}$, and

$$I_+[r] \triangleq \begin{cases} 0 & \text{if} \quad r \geqslant 0 \\ \infty & \text{if} \quad r < 0 \end{cases}$$

is the indicator function for the positive reals. The system in Eq. (52) has no constraints. However, we cannot apply the Newton–Raphson method, since the left-hand side of Eq. (52) is not differentiable with respect to $\mathbf{u}$. The main idea of the barrier method is to approximate the indicator function I_+ by the logarithmic function $\hat{I}_+[r]=-\ln(r)/\lambda$, where $\lambda>0$ is a parameter that controls the accuracy of the approximation. As λ increases, the approximation of $I_+[r]$ by $\hat{I}_+[r]$ becomes more accurate. Substituting I_+ by $\hat{I}_+$ in Eq. (52), we obtain the following system of equations:

$$\mathbf{f}(\mathbf{u}) \triangleq \mathbb{C}\mathbf{h}(\mathbf{u}) - \frac{1}{\lambda}\sum_{m\in\mathcal{M}} \ln[e_m(\mathbf{u})] = \mathbf{0}. \tag{53}$$

Function $\mathbf{f}$ in Eq. (53) is now differentiable. Therefore, the Newton–Raphson method can be applied to find a vector $\hat{\mathbf{u}}$ that satisfies Eq. (53). Of course, $\hat{\mathbf{u}}$ will only be an approximate solution to Eq. (52). Intuition suggests that the quality of approximation will improve for larger values of parameter λ.[37]

Application of the previous method for approximating fast reaction kinetics in complex stochastic biochemical systems requires thorough investigation of several theoretical and practical issues (e.g., does Eq. (51) has a solution, is the solution unique, how fast the Newton–Raphson method converges, etc.). This investigation is beyond the scope of the present paper, whose main objective is to analytically illustrate the feasibility of employing the SQEA for eliminating fast reaction kinetics. Details will become available in a future contribution.

To conclude, we note that the work presented in this paper was developed independently from a recently published work on the same subject by Cao *et al.*[38] These investigators present a systematic approach to the problem of approximating fast reaction kinetics in stochastic biochemical systems that clearly explains the basic issues and challenges underlying such an approximation. Our approach is closely related to theirs, with several similarities and differences, which we briefly discuss below.

Cao *et al.* model the state of a stochastic biochemical system by the Markovian molecular population vector $\mathbf{X}(t)$. This is in contrast to our method, which uses the DAs to characterize the state of the system. As we briefly discussed in Sec. II, the DAs may provide a more appropriate characterization of the state of a stochastic biochemical system with delays.

By defining a slow species to be any species whose population does not get changed by *any* fast reaction, and a fast species to be any species whose population gets changed by *some* fast reaction, Cao *et al.* partition the state vector $\mathbf{X}(t)$ into slow and fast subvectors $\mathbf{X}^s(t)$ and $\mathbf{X}^f(t)$, respectively. They note that $\mathbf{X}^f(t)$ is a non-Markovian process and, therefore, it cannot be characterized by a CME. To ameliorate this problem, they switch off the slow reaction subsystem and introduce a *virtual* fast process $\hat{\mathbf{X}}^f(t)$, composed of the same fast state variables as $\mathbf{X}^f(t)$, which evolves only through the fast reactions. It turns out that $\hat{\mathbf{X}}^f(t)$ is a Markov process, characterized by the CME (5) in Cao *et al.* This is similar to the step that leads to the CME (11) in our paper.

By assuming that within two consecutive occurrences of slow reactions, the virtual fast process rapidly reaches steady state, Cao *et al.* determine an "effective" propensity function for a slow reaction whose product with a small time interval dt provides an acceptable approximation of the probability that the reaction will occur within the time interval $[t,t+dt)$. By using a set of (rather complicated) arguments, they show that this propensity function is the mean of the actual

propensity function of the slow reaction with respect to the steady-state probability distribution of the virtual fast process, given by Eq. (9) in Ref. 38. In contrast to these steps, we have shown in our paper that the slow reaction subsystem is *exactly* characterized by the CME (9), with propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, given by Eq. (10). Equation (10) is the analog of Eq. (9) in Ref. 38.

Cao *et al.* show that computation of the effective propensity functions requires only knowledge of the first two moments of the steady-state distribution of the virtual fast process. This is similar to our conclusion that computation of the propensity functions $\alpha_m^{(t)}(\mathbf{z}_s)$, $m \in \mathcal{M}_s$, given by Eq. (10), requires only knowledge of the conditional mean and covariance matrix of the fast DA process $\mathbf{Z}_f$.

The main difference between the work of Cao *et al.* and ours is that the former is based on a statistical version of the well-known quasisteady-state assumption (QSSA) of chemical kinetics, whereas, the latter is based on a statistical version of the well-known SQEA. Note that the QSSA does not necessarily imply the SQEA, and vice versa. As a matter of fact, the fast reactions may reach a state of equilibrium before the fast species reach steady state. In addition to the fact that the QSSA can only be applied to biochemical systems with stable fast reactions, it requires that the time between consecutive occurrences of slow reactions is large enough so that, during that time, the fast reaction subsystem reaches steady state. The SQEA simply assumes that, within consecutive occurrences of slow reactions, the fast reactions subsist nearly at equilibrium. This is a common assumption for approximating the dynamics of a biochemical system consisting of fast reactions after a short initial relaxation period.[15]

[1] H. H. McAdams and A. Arkin, Trends Genet. **15**, 65 (1999).
[2] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).
[3] D. T. Gillespie, Physica A **188**, 404 (1992).
[4] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[5] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).
[6] M. A. Gibson and J. Bruck, J. Phys. Chem. A **104**, 1876 (2000).
[7] D. T. Gillespie, J. Chem. Phys. **115**, 1716 (2001).
[8] D. T. Gillespie and L. R. Petzold, J. Chem. Phys. **119**, 8229 (2003).
[9] M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie, J. Chem. Phys. **119**, 12784 (2003).
[10] K. Burrage, T. Tian, and P. Burrage, Prog. Biophys. Mol. Biol. **85**, 217 (2004).
[11] Y. Cao, H. Li, and L. Petzold, J. Chem. Phys. **121**, 4059 (2004).
[12] T. Tian and K. Burrage, J. Chem. Phys. **121**, 10356 (2004).
[13] A. Chatterjee, D. G. Vlachos, and M. A. Katsoulakis, J. Chem. Phys. **122**, 024112 (2005).
[14] C. V. Rao and A. P. Arkin, J. Chem. Phys. **118**, 4999 (2003).
[15] R. Heinrich and S. Schuster, *The Regulation of Cellular Systems* (Chapman and Hall, New York, 1996).
[16] E. L. Haseltine and J. B. Rawlings, J. Chem. Phys. **117**, 6959 (2002).
[17] M. A. Savageau, *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology* (Addison-Wesley, Reading, MA, 1976).
[18] D. T. Gillespie, J. Chem. Phys. **113**, 297 (2000).
[19] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1992).
[20] Haseltine and Rawlings (Ref. 16) refer to $Z_m(t)$ as the *extent of reaction*, a term that is already used to characterize the change in the number of molecules of a reactant or product species caused by a given reaction, divided by the stoichiometric coefficient associated with that species (Ref. 15). To avoid confusion, we prefer to use here the term *degree of advancement* (Ref. 19).
[21] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. (Academic, San Diego, 1975).
[22] R. Srivastava, L. You, J. Summers, and J. Yin, J. Theor. Biol. **218**, 309 (2002).
[23] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, S. L. Zipursky, and J. Darnell, *Molecular Cell Biology*, 5th ed. (Freeman, New York, 2003).
[24] Parameters $c_2, c_3, \ldots, c_6$, have been obtained from the data published by Srivastava and co-workers (Ref. 22). From these data, we have determined that $c_1 = 1.5 \times k_4$, where $k_4 = 7.5 \times 10^{-6}$ day^{-1}. Therefore, we set $c_1 = 11.25 \times 10^{-6}$ day^{-1}.
[25] G. K. Ackers, A. D. Johnson, and M. A. Shea, Proc. Natl. Acad. Sci. U.S.A. **79**, 1129 (1982).
[26] M. A. Shea and G. K. Ackers, J. Mol. Biol. **181**, 211 (1985).
[27] A. Arkin, J. Ross, and H. H. McAdams, Genetics **149**, 1633 (1998).
[28] J. Hasty, J. Pradines, M. Dolnik, and J. J. Collins, Proc. Natl. Acad. Sci. U.S.A. **97**, 2075 (2000).
[29] M. Ptashne, *A Genetic Switch: Phage Lambda Revisited*, 3rd ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2004).
[30] R. Bundschun, F. Hayot, and C. Jayaprakash, Biophys. J. **84**, 1606 (2003).
[31] M. Santillán and C. Mackey, Biophys. J. **86**, 75 (2004).
[32] P. J. Darling, J. M. Holt, and G. K. Ackers, J. Mol. Biol. **302**, 625 (2000).
[33] F. J. Isaacs, J. Hasty, C. R. Cantor, and J. J. Collins, Proc. Natl. Acad. Sci. U.S.A. **100**, 7714 (2003).
[34] H. Bremer and P. P. Dennis, in *Escherichia coli and Salmonella thyphymurium: Cellular and Molecular Biology*, edited by F. C. Neidhart, R. Curtis, J. L. Ingraham *et al.* (American Society for Microbiology, Washington, 1996), Vol. 2, pp. 1553–1569.
[35] T. Lu, D. Volfson, L. Tsimring, and J. Hasty, Syst. Biol. **1**, 121 (2004).
[36] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. (Cambridge University Press, Cambridge, UK, 1993).
[37] S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, UK, 2004).
[38] Y. Cao, D. T. Gillespie, and L. R. Petzold, J. Chem. Phys. **122**, 014116 (2005).