Perfect Sampling of the Master Equation for Gene Regulatory Networks

Martin Hemberg and Mauricio Barahona
Department of Bioengineering and Institute for Mathematical Sciences, Imperial College London, London, United Kingdom

ABSTRACT We present a perfect sampling algorithm that can be applied to the master equation of gene regulatory networks. The method recasts Gillespie’s stochastic simulation algorithm (SSA) in the light of Markov chain Monte Carlo methods and combines it with the dominated coupling from the past (DCFTP) algorithm to provide guaranteed sampling from the stationary distribution. We show how the DCFTP-SSA can be generically applied to genetic networks with feedback formed by the interconnection of linear enzymatic reactions and nonlinear Monod- and Hill-type elements. We establish rigorous bounds on the error and convergence of the DCFTP-SSA, as compared to the standard SSA, through a set of increasingly complex examples. Once the building blocks for gene regulatory networks have been introduced, the algorithm is applied to study properly averaged dynamic properties of two experimentally relevant genetic networks: the toggle switch, a two-dimensional bistable system; and the repressilator, a six-dimensional transcriptional oscillator.

INTRODUCTION

The recent interest in stochastic models of chemical reactions has been largely motivated by experiments that have demonstrated the stochastic nature of key processes in the cell, such as signaling or gene regulation (1–10). The inherent stochasticity of these biochemical processes is due to the low number of molecules involved in the reactions (11), amplified in some cases by the proximity to a critical point of the dynamics of the system (12). Low copy numbers lead to intrinsic fluctuations and to the breakdown of models based on differential equations (13). A number of stochastic models of gene regulation (11,14–21) and signaling (22–24) have been formulated and analyzed numerically with the aim of identifying the sources of randomness, and how noise is controlled and harnessed in the cellular environment.

The starting point for such stochastic descriptions is the master equation (ME), a conservation equation that gives the probability of finding the system in state \( i \) at time \( t \),

\[
P_i(t) = P_i(t) = \sum_j W_{ij} P_j(t) - W_{ji} P_i(t),
\]

where \( W_{ij} \) is the transition rate from state \( i \) to state \( j \). The application of the ME to chemical systems goes back at least to the study of an irreversible reaction by Delbrück (25). This work has been extended by a number of authors (26,27). Equation 1 is usually referred to as the chemical master equation (CME) when it describes stochastic versions of law-of-mass-action systems. In this article, we extend the notation and apply the term CME to compound reaction rates (e.g., Michaelis-Menten) as well.

Although theoretically rigorous, there are very few systems for which the ME has been solved analytically. Direct numerical integration is often very difficult, due to the fact that the state space grows very rapidly, and also to the stiffness of the equations. One way to overcome these issues is to employ some kind of approximation scheme. In certain limits, one can consider continuum approximations leading to partial differential equations, such as van Kampen’s linear-noise approximation and Fokker-Planck equations (13). However, these approximations disregard the discreteness of the state of the system, and can therefore give rise to significant deviations when the number of molecules is very small, as can be the case for gene regulatory networks (GRNs).

A different approach for dealing with the CME numerically is provided by the widely used stochastic simulation algorithm (SSA) by Gillespie (28). The SSA is a kinetic Monte Carlo algorithm, rigorously derived from the same assumptions as the CME, which gives realizations of the trajectory of a given CME. The SSA follows an idea that can be traced back to Doob (29) and provides an exact procedure for the kinetic sampling of the CME (30) in the sense that we obtain an unbiased and convergent estimate of the solution of the CME. Due to its biological applications, there has been considerable interest in the SSA in recent years and several extensions have been proposed (e.g., the explicit introduction of space (31,32) and time-delays (14)), as well as algorithmic speed-up improvements (33,34).

In many experiments of interest (3,7,9,29), and in related theoretical analyses (11,16,18,21–23,35,36), the system at hand is assumed to have reached the stationary distribution, i.e., the left-hand side of Eq. 1 is zero. Consider the ME (Eq. 1) in operator form,

\[
\dot{P} = WP - \text{diag}(e^T W) P = Q^T P,
\]

where \( P \) is a (possibly infinite-dimensional) column vector of probabilities, \( e \) is the vector of ones, and \( W \) is the transition matrix.
matrix. The stationary distribution $\mathbf{P}^*$ (usually denoted $\pi$ in Markov theory) could in principle be obtained from the first left eigenvector of the $Q$-matrix (37). However, this approach is impractical due to the curse of dimensionality: the run-time and memory requirements scale exponentially with the number of types of molecules. For small state spaces, one can still consider a (truncated) finite subspace to obtain an approximate convergent solution. Our own accurate version of this approximate eigenvector method is used in this article to check the accuracy of our DCFTP-SSA algorithm when an analytical expression for the stationary distribution is not known.

To sample from the stationary solution of Eq. 1, one would have to run the SSA for an infinite time. Of course, the most common practical solution is to run the algorithm repeatedly for a very long time and hope that the system has reached the stationary distribution when the run is stopped (15). The lack of a termination certificate can lead to computational inefficiency if the runs are longer than necessary or, more importantly, to missampling from the wrong distribution if the runs are not long enough.

In this article, we present an algorithm (DCFTP-SSA) that guarantees perfect sampling of the stationary solution of the CME for the general class of biochemical networks formed by the interconnection of generalized Hill- and Monod-type reactions, the canonical models for GRNs and enzymatic networks. The DCFTP-SSA builds on the standard SSA and considers it in the light of Markov chain Monte Carlo algorithms, specifically within the dominated-coupling-from-the-past (DCFTP) framework introduced by Wilson (38) and Kendall (39). Henceforth, we introduce the algorithm and its properties in detail through the analysis of the building blocks of GRNs. We then apply the DCFTP-SSA to the study of two systems of experimental interest in synthetic biology: the toggle switch (6) and the repressilator (8).

THE DOMINATED COUPLING FROM THE PAST STOCHASTIC SIMULATION ALGORITHM (DCFTP-SSA)

In many applications, we are interested in sampling from a complicated distribution that cannot be written down explicitly. Markov chain Monte Carlo methods can sometimes provide an answer by setting up a Markov chain, which has the desired distribution as its stationary distribution. Sampling from the target distribution is then achieved by evolving the Markov chain until it has reached its stationary distribution (37). The major issue in these schemes is when to stop the Markov chain. This problem was addressed by Propp and Wilson with their coupling from the past (CFTP) algorithm (38), a celebrated example of what are commonly referred to as perfect sampling algorithms (40). The version used here is the dominated coupling from the past (DCFTP) algorithm, the extension introduced by Kendall (39) to study continuous-time Markov processes on an unbounded state space. We now give a brief introduction to the algorithm.

In principle, a Markov chain would have to be run for an infinite time to reach its stationary distribution. The CFTP algorithm, however, recasts the problem as a procedure in which the running time becomes a random variable but a certificate is issued if the stationary distribution has been reached. This is signaled by the coalescence of relevant coupled Markov chains. Markov chains are coupled if they have the same transition rules and use the same sequence of random numbers for their realization. Coupled Markov chains are said to coalesce when they meet. Clearly, coupled chains with the same transition rule (but started from different initial states) will have the same values for all future times after the coalescence time $T_c$. Propp and Wilson proposed to use Markov chains coupled from the past to ensure that the whole state space of initial conditions maps to the same state at present. By extending sufficiently far back into the past and using a fixed sequence of random numbers, we will eventually reach a time from which all paths map to the same state at $t = 0$. This condition is equivalent to the stationary distribution, since the starting condition is irrelevant for the current state.

In many systems, the state space is very large, making it infeasible to monitor all paths and their coalescence. However, the situation is much simpler if the state space is partially ordered and the partial ordering is maintained by the transition rules. This is the case if we have a monotone transition rule $\phi$, such that $\phi(x, R) \leq \phi(y, R)$ if $x \leq y$, where $\leq$ denotes the partial ordering, $x$ and $y$ are states and $R$ is a random number used to determine the transition. The algorithm also works for anti-monotone transition rules: $\phi(x, R) \geq \phi(y, R)$ if $x \leq y$ (41). For a state space to be partially ordered, it must be reflexive, antisymmetric, and transitive. If the monotonicity and partial ordering hold, we only need to monitor two paths: an upper path $U$ and a lower path $L$. Since all other paths lie in between and will coalesce when $U$ and $L$ have done so. This sandwiching property is illustrated in Fig. 1. The guaranteed sampling of the CFTP scheme comes with a tradeoff: the run time is unbounded, since $T_c$ is itself a random variable, but it is finite almost surely. If the run is prematurely terminated by an impatient user, the sample will be biased.

If the system has an unbounded state space, we must instead employ Kendall’s dominated CFTP algorithm (39). The DCFTP relies on finding a reversible dominating process $D$, which bounds the original process from above and for which the stationary distribution is known. To create a dominating process, we exploit the reversibility and stationarity of $D$: start the chain $\tilde{D}$ at $t = 0$ from the stationary distribution and evolve it until $t = T$; then use $D_{i-1} = \tilde{D}_{i}$ as the dominating process on the interval $[-T, 0]$. It then follows that all chains of the original dominated process started at $t = -\infty$ will be less than or equal to $D$ at time $-T$ and, consequently, chains coupled to $D$ started from a state $U_{-T} \leq D_{-T}$ can be interpreted as random realizations of the original (dominated) process started from $t = -\infty$. 
A brief outline of the DCFTP-SSA is as follows:

1. Run the dominating process \( \tilde{D} \) with known stationary distribution forward in time from \( t = 0 \) until \( t = T \).
2. Apply time reversal to the stationary process \( \tilde{D} \) to obtain the dominating process \( D \) such that \( D_t = \tilde{D}_{-t} \) from \( t = -T \) to \( t = 0 \).
3. Start upper (\( U \)) and lower (\( L \)) chains of the original process starting from \( U_{-T} = D_{-T} \) and \( L_{-T} = 0 \) and update each chain coupled to \( D \) until \( t = 0 \).
4. If the chains have coalesced at \( t = 0 \), the common value \( U_0^* = L_0^* \) is a sample from the target distribution \( P^* \).
5. If coalescence has not occurred, double the running time to \( t = 2T \) reusing the random numbers from the previous iteration and repeat.
6. Keep doubling the running time until coalescence has occurred at \( t = 0 \).

The main feature of the DCFTP-SSA is that it provides a certificate for correct sampling from the stationary distribution. Therefore, we can use the DCFTP-SSA to sample, numerically, the stationary distribution of the system with guaranteed Monte Carlo accuracy. This can be of importance in high-dimensional systems with complicated landscapes where the SSA simulations can converge slowly. In the first two sections below, we test the application of the DCFTP-SSA to the sampling of stationary distributions. In those cases, we only use the final value of each run as a sample, thus ensuring that the samples are independent.

In addition, we can use the algorithm to discard the burn-in period in stochastic simulations by running the DCFTP-SSA until a guaranteed sample from the stationary distribution is obtained and then using it to initiate a run of the ordinary SSA from time \( t = 0 \) onwards. This is important for the numerical study of stationary properties of systems with high variability, e.g., with underlying oscillatory or excitatory behavior. In the final section, we present examples of these applications to the toggle switch and the repressilator.

**Detailed application to the first-order reaction**

To illustrate the DCFTP-SSA, we study in detail the simple one-dimensional first-order reaction, for which a full time-dependent analytical solution of the CME is available. This allows us to study the convergence of the algorithm, as compared to the standard SSA, and to explore the difference between two sources of error, which often get entangled: finite sampling error and the missampling error due to the fact that we may not have reached the true stationary distribution.

Consider the simple first-order reaction, which has been used as a very simplified description of transcription and translation (2,21),

\[
\emptyset \xrightarrow{k_1} A \xrightarrow{k_2} \emptyset \Rightarrow A = k - A,
\tag{3}
\]
where species $A$ is created at a (normalized) constant rate $k$ from a source ($\delta$) and degraded to a sink ($\emptyset$). The corresponding CME is

$$\dot{P}_j = kP_{j-1} - kP_j + (j+1)P_{j+1} - jP_j = (E^{-1} - 1)kP_j + (E - 1)jP_j,$$  

(4)

with $P_j$ denoting the probability of having $j$ molecules of $A$. $E^{-1}$ and $E$ are the step operators defined by van Kampen (31): $Ef(j) = fj + 1$ and $E^{-1}f(j) = f(j - 1)$ for a function $f(j)$.

Equation 4 is one of a few CMEs for which the analytical expression of the stationary solution is known (13,42): it is the Poisson distribution with parameter $\lambda = k$. More importantly for our purposes, one can obtain the full time-dependent solution of the CME (see Appendix 1). For the usual initial condition with 0 molecules, the time-dependent distribution $P(t)$ is given by (see Fig. 2, inset)

$$P_j(t) = P(j,t|0,0) = \frac{\exp(-k(1 - e^{-t}))}{j!}[k(1 - e^{-t})]^j,$$  

(5)

which converges to the correct stationary distribution as $t \to \infty$.

The missampling error can be understood analytically in this simple example. If the standard SSA is used to estimate the stationary solution of the CME (Eq. 4), $N$ samples from independent SSA runs will be collected at a stopping time $t_{SSA}$. This leads to the sampled distribution $P_{SSA}(N, t_{SSA})$, which clearly converges to the stationary distribution $P^*$ in the double limit $t_{SSA}, N \to \infty$. As $N$ is increased, with fixed $t_{SSA}$, we sample with increasing Monte Carlo accuracy from $P(t_{SSA})$ given by Eq. 5 but not from the true stationary distribution $P^*$. We therefore reach an error floor that cannot be broken. Using the analytical expression (Eq. 5), the asymptotic error floor $\epsilon_E(t_{SSA})$ is shown to be

$$\epsilon_E(t_{SSA})^2 = \sum_{j=0}^{\infty} (P_j - P_j(t_{SSA}))^2 = \sum_{j=0}^{\infty} \left( \frac{e^{-k}k^j}{j!} - \frac{e^{-\text{ss}_{SSA}}(ak)^j}{j!} \right)^2$$.  

where $\text{ss} = 1 - e^{-\text{ss}_{SSA}}$ and $I_0(\alpha)$ is the modified Bessel function of the first kind.

This source of error is eliminated in a DCFTP-SSA formulation of this process. The key step is to find a reversible dominating process with known stationary distribution. In this example, it is clear that the best choice is to use the process (Eq. 4) itself, i.e., the $U$ and $D$ chains in Fig. 1 will be identical. Fig. 2 shows the Euclidean error for the DCFTP-SSA sample distribution, which shows no flooring and the expected $N^{-1/2}$ scaling with the number of Monte Carlo runs (43). Equivalent results are obtained with other error measures for the sampled distribution, e.g., the Kullback-Leibler and Kolmogorov distances, and the $\chi^2$-goodness of fit.

THE BUILDING BLOCKS: NETWORKS AND NONLINEAR ELEMENTS

To extend our study to GRNs, we need to consider how to deal with two key ingredients: networks of reactions, and nonlinear elements of the Hill and Monod types.

Application to networks: simple gene model of two coupled first-order reactions

To showcase how to apply the DCFTP-SSA to networks of several species, we use a widely studied, simplified model of gene expression (17,19,21). In its simplest form, the model consists of two types of molecules: mRNA ($M$) and proteins ($R$), which are produced and depleted at constant rates. In addition, the mRNA catalyzes protein production through a linear enzymatic reaction:

\[
\begin{align*}
\emptyset & \xrightarrow{k_M} M \\
\emptyset & \xrightarrow{k_B} R \\
M & \xrightarrow{k_R} R
\end{align*}
\]

The deterministic description is given by...
\[ \dot{M} = k_M - M \]

and the corresponding CME is
\[ \dot{P}_{M,R} = (E_M^{-1} - 1)k_M P_{M,R} + (E_M - 1)M P_{M,R} \]
\[ + (E_R^{-1} - 1)(k_R + k_RM - R). \]  

In fact, Eq. 8 has been solved: the stationary state of the CME of the system, indicating the need for longer SSA runs to avoid missampling from the true distribution. The DCFTP-SSA depends on the dimensionality of the system. Fig. 3 (inset) shows that the marginal distribution (for \( M \)) converges like the one-dimensional reaction in Fig. 2, with similar error floors for the standard SSA. Note, however, that the joint distribution (Fig. 3) exhibits slower convergence of the standard SSA due to the higher dimensionality of the system, indicating the need for longer SSA runs to avoid missampling from the true distribution. The DCFTP-SSA shows \( N^{-1/2} \) Monte Carlo scaling with no error floors regardless of the dimensionality of the system.

**Nonlinear elements: Hill and Monod reactions**

The Hill reaction is widely used to model the sigmoidal (nonlinear) characteristics of many biological processes and specifically those involved in genetic regulation (4,14,24). This model incorporates negative feedback, whereby the rate of creation of new molecules decreases as their concentration increases:

\[ \theta \rightarrow f(A,k,\alpha) \rightarrow \emptyset \Rightarrow \dot{A} = \frac{k}{\theta^\alpha + A^\alpha} - A. \]  

Here, \( f(A;k, \alpha) = k/(\theta^\alpha + A^\alpha) \) is the state-dependent reaction rate, \( k \) is the renormalized reaction constant, and \( \alpha \) is the cooperativity factor (44). Equation 9 can be derived from elementary law-of-mass-action kinetics through elimination of variables (18). The corresponding CME is given by

\[ \dot{P}_j = (E_j^{-1} - 1)\frac{k}{\theta^\alpha + j} P_j + (E_j - 1)P_j. \]  

It is usual to fix \( \theta = 1 \) and we do so in the following. For the particular case \( \alpha = 1 \), Eq. 9 reduces to the familiar Michaelis-Menten equation, for which we can obtain the following analytical expression for the stationary distribution (see Appendix B): \( P_{\infty} = ck^j/j! \), where \( c = 1/I_0(2\sqrt{K}) \) is a normalization constant.

The Monod reaction is used to model gene upregulation, as referred to in particular to the autocatalysis common in many eukaryotes (3,45,46). The reaction can be written as

\[ \theta \rightarrow g(A,k,\alpha) \rightarrow \emptyset \Rightarrow \dot{A} = \frac{ka^n}{\theta^\alpha + A^\alpha} - A, \]  

where \( g(A;k, \alpha) = ka^n/(\theta^\alpha + A^\alpha) \) is the state-dependent reaction rate that encapsulates the positive feedback. The corresponding CME is given by

\[ \dot{P}_j = (E_j^{-1} - 1)\frac{k j^n}{\theta^\alpha + j} P_j + (E_j - 1)P_j. \]  

Again, we fix \( \theta = 1 \) in the following. The stationary distribution of the Monod CME (Eq. 12) with \( \alpha = 1 \) can be obtained analytically (see Appendix B): \( P_{\infty} = ck^j/j! \), with normalization constant \( c = 1/\Gamma(2k - \alpha) \).

For general \( \alpha \), no explicit solution for the stationary distribution of the CMEs Eq. 10 and Eq. 12 is known, and numerical methods are therefore necessary. Fortunately, the DCFTP-SSA is directly applicable to the reversible Hill and Monod CMEs: the first-order reaction, studied in detail in the previous section, can be used to produce a dominating process for both processes. To see this, note that Eq. 10 and Eq. 12 have birth rates that are bounded from above by \( k \) at all times. It is thus straightforward to make sure that the upper and lower paths have rates that fulfill the requirements of the algorithm. As an illustration, Fig. 1 shows one DCFTP-SSA run for the Hill CME (Eq. 10), where \( D \) corresponds to the dominating first-order linear process (Eq. 4) and \( U \) and \( L \) are the coupled Hill processes (39,40). Our detailed simulations of the Hill and Monod CMEs (data not shown) show \( N^{-1/2} \) Monte Carlo scaling.
convergence of the DCFTP-SSA unaffected by the missampling errors that can appear when using the standard SSA with short stopping times.

APPLICATION TO NONLINEAR MODELS OF GENE REGULATORY NETWORKS

Our foregoing discussion clarifies why the DCFTP-SSA is applicable to reversible GRNs consisting of interconnected linear enzymatic, Hill and Monod birth reactions: it is possible to define a dominating process for these networks based on the associated linear network for which the stationary solution (a multivariate Poisson distribution) is known. We now apply our scheme to two recent, canonical examples of synthetic GRNs: a bistable genetic toggle switch (8), with a bimodal stationary distribution; and the repressilator (6), a simple synthetic oscillator.

Toggle switch: two coupled Hill equations

An important characteristic of biochemical networks is the possibility of multistability (1,8,15,16,18,21,47,48), as exemplified by the toggle switch designed by Gardner et al. by combining two mutually repressing genes (8,11,15,16,18). This leads to a bistable system with a sharp switching threshold:

\[
\begin{align*}
&f(B; k_A, \alpha) = 1 \\
&f(A; k_B, \beta) = 1
\end{align*}
\]

A deterministic model can be derived using two coupled Hill equations:

\[
\begin{align*}
\dot{A} &= \frac{k_A}{1 + B^\beta} - A \\
\dot{B} &= \frac{k_B}{1 + A^\beta} - B,
\end{align*}
\]

where \(A\) and \(B\) are the molecular numbers of the transcription factors. With appropriate cooperativity factors \(\alpha, \beta > 1\) and reaction constants \(k_A\) and \(k_B\), Eq. 13 has two stable fixed points (8).

The corresponding CME is given by

\[
\begin{align*}
\dot{P}_{A,B} = (E_A^{-1} - 1) \frac{k_A}{1 + B^\beta} P_{A,B} + (E_A - 1)AP_{A,B} \\
&\quad + (E_B^{-1} - 1) \frac{k_B}{1 + A^\beta} P_{A,B} + (E_B - 1)BP_{A,B}.
\end{align*}
\]

Based on our discussion above, it is easy to see that linear process with constant birth rates \(k_A\) and \(k_B\) will be a dominating process for Eq. 14. Since the Hill function is antimonotone, we must use a crossover scheme for the updates to make sure that all initial conditions are sandwiched between the upper and lower paths, as described in Häggström and Nelander (41). We have applied the DCFTP-SSA to Eq. 14 to obtain the stationary distribution of the system, which is clearly bimodal (Fig. 4). Because there is no analytical solution in this case, we have checked the accuracy and convergence of the DCFTP-SSA against the approximate eigenvector method with excellent results (not shown).

Beyond certifying proper convergence, the DCFTP-SSA can be used to study properties of the network that need to be properly averaged over the stationary distribution. Consider the mean time for the system to switch from the basin of one fixed point \(A^#\) to the other \(B^#\), i.e., the escape time of the CME (Eq. 14). If the standard SSA were to be used, it would be unclear how to collect proper Monte Carlo statistics, since no certificate of stationarity exists. In contrast, we use a mixed scheme in which an initial DCFTP-SSA run is followed by a (faster) SSA run. The initial DCFTP-SSA run ensures that the initial condition for the SSA run is correctly sampled from the stationary distribution at \(t = 0\), in essence eliminating the burn-in period. From that point onwards, a standard SSA run is performed to obtain statistics of the average time to cross the separators. The results are reported in Fig. 4.

In their experiments, Gardner et al. (21) were able to induce faster \(A^# \rightarrow B^#\) switching by reducing the down-regulation capability of \(A\) in the cellular environment. This can be mimicked by a modified CME in which the \(A^#\)-term in Eq. 14 is removed. If we run the SSA of this modified CME starting from the stationary distribution of the original system (Eq. 14), the escape time \(\tau_{A^# \rightarrow B^#}\) becomes approximately

\[
\tau_{A^# \rightarrow B^#} = 67.3 \pm 2.4 (65.9)
\]

\[
\tau_{A^# \rightarrow B^#} = 68.6 \pm 2.6 (70.3)
\]

FIGURE 4 Stationary distribution for the toggle switch (Eq. 14) with parameters \(k_A = 30, k_B = 10, \alpha = 3, \beta = 1\). The bimodality of the distribution leads to switching behavior. We have checked that the Euclidean error of the DCFTP-SSA sampled distribution decreases at the expected \(N^{-1/2}\) rate (data not shown). The state space can be divided in two regions \(A^#\) and \(B^#\) corresponding to the two basins of attraction of the fixed points. The separatrix is found from the sign of the Fiedler eigenvector of the Laplacian (47) of the state space lattice. The probabilities \(P^a_A\) and \(P^b_B\) of finding the system in each basin computed with the DCFTP-SSA compare very well with those obtained with the approximate eigenvector method (in parenthesis). Similarly, the expected time to reach the separators averaged over all the states in each basin (37) calculated with the DCFTP-SSA (averaged over 25 ensembles, each consisting of 1000 samples) compares well with the approximate eigenvector method (in parenthesis).
two orders-of-magnitude smaller, in broad agreement with the experiment. In fact, it can be shown that the stationary distribution of the modified CME becomes unimodal, with no observable peak for species A.

Repressilator: six-dimensional system of linear enzymatic and Hill equations

The repressilator, a synthetic system of transcriptional regulators, is perhaps the simplest biochemical oscillator that has been implemented experimentally (6). It consists of three genes in a loop, in which the expression of one gene is inhibited by the product of another gene in succession (Fig. 5).

The repressilator is the product of another gene in succession (Fig. 5).

\[ \dot{M}_i = \frac{k_M}{1 + R^n_{i+1}} - d_M M_i, \]
\[ \dot{R}_i = k_R + k_R M_i - R_i, \quad i = 0, 1, 2 \pmod{3}. \]  

(15)

Here \( M_i \) values are the mRNA levels (with production rate \( k_M \) and degradation rate \( d_M \)) and \( R_i \) values are the corresponding proteins (with basal rate \( k_R \) and linear production rate \( k_P \)). This model of the repressilator is therefore a network of linear enzymatic elements (Eq. 8) for the protein production terms and Hill reactions (Eq. 10) to modulate the production of mRNA. The deterministic system (Eq. 15) has been shown to be oscillatory, with the concentrations of the three proteins peaking in succession (Fig. 5, bottom panel).

The corresponding CME is given by

\[ \dot{P}_n = \sum_{i=0}^{2} (\mathbb{E}_{E_{i+1}} - 1) \cdot \frac{k_M}{1 + R^n_{i+1}} P_n + (\mathbb{E}_{E_i} - 1) d_M M_i P_n + (\mathbb{E}_{E_i} - 1) k_R M_i P_n + (\mathbb{E}_{E_i} - 1) k_R P_n. \]

(16)

where the shorthand \( P_n \) denotes the state \( P_{M_0, M_1, M_2, R_0, R_1, R_2} \) and \( i = 0 \) mod 3. We have simulated this six-dimensional system using the crossover scheme described above: we use the DCFTP-SSA to discard the burn-in period, i.e., the time it takes for the Markov process to reach the stationary distribution, which is the initial conditions for the SSA runs are sampled from the stationary distribution. As a measure of the stationarity of our DCFTP-SSA initialized sampling, we have checked its ergodicity by establishing through an F-test that the average period and amplitude of several short simulations are statistically indistinguishable from those obtained from one long simulation.

As can be seen in the sample trajectory in Fig. 5, the stochastic system is extremely noisy but it retains the overall oscillation of the genes in succession. The oscillations do not die out even for very long simulations and we have collected statistics of such long runs. Fig. 5 shows a robust feature of the oscillator: the distribution of the period is well fitted by a generalized Gamma distribution, which is characteristic of excitable systems with a refractory period. The oscillatory behavior is present for a wide range of the parameters \( k_M, k_R, \) and \( d_M \), which could be modified experimentally. Fig. 5 summarizes our investigation of the robustness of the repressilator to parametric variation. Note the good agreement of the calculated mean period with the corresponding deterministic values. The simulations show that changes in parameters \( k_M \) and \( k_R \) produce a similar, small effect on the period. The system is, however, more sensitive to parametric changes in \( d_M \). The evaluation of how these parametric variations could be used for the design of tunable and reliable oscillators will be the subject of further study (49).

DISCUSSION

In recent years, the importance of stochastic effects in gene regulation has been elegantly demonstrated through experiments. From a theoretical perspective, such systems are usually analyzed numerically with the standard SSA. When studying stationary properties, this raises questions about the sampling of the stochastic process since the SSA does not provide explicit guarantees for convergence. One issues of interest for the use of the DCFTP-SSA is that the shape of the stationary distribution can be more important than the dimensionality of the system for the convergence of the stochastic simulations. For systems with smooth unimodal distributions, such as the repressilator, the SSA approaches the stationary distribution rapidly, regardless of the initial conditions. However, the ordinary SSA can be highly sensitive to initial conditions for systems with multimodal distributions, such as the toggle switch. In general, the SSA will converge rapidly to the nearest mode. If we are primarily interested in escape times, they can be estimated accurately by using the ordinary SSA started close to one of the modes. However, if the escape times from the modes are long and the ordinary SSA is used, we run the risk of missampling the stationary distribution, specifically in important areas of low probability such as the separators. The DCFTP-SSA circumvents this problem but at the cost of longer coalescence times.

The guaranteed sampling provided by the DCFTP-SSA comes at an extra computational cost. In general, the coalescence time will depend on the topology of the network and the form of the reactions. However, the CPU overhead is in no way prohibitive and the DCFTP-SSA can be run on an ordinary desktop computer for the systems presented in this article. An important theoretical feature of the DCFTP-SSA is the fact that its runtime is not bounded. As our simulations show, this feature does not seem to impose practical restrictions on the algorithm. If necessary, however, this issue can be resolved by using FMMR (50), another perfect sampling algorithm, which dispenses with variable stopping times by running the Markov chains for a fixed time and using rejection sampling to discard those simulations that have not reached the stationary distribution.

In its current form, the DCFTP-SSA can be applied to reversible systems of linear and nonlinear (anti-)monotonic unimolecular reactions for which there exists a dominating process with known stationary distribution. We note that...
the algorithm is applicable not only to networks of linear enzymatic, Hill and Monod birth reactions of standard form, but also to reactions whose rate equation is a rational function that can be dominated by a linear process. This condition is equivalent to requiring that the polynomial in the denominator of the propensity function have a higher degree than the polynomial in the numerator. These types of rational functions are frequently obtained as compound unimolecular rate laws to represent more complex enzymatic (45) or gene regulatory reactions (51). An example of such functions is the model for the \( \lambda \)-phage lysis-lysogeny switch (11,16), which we have also simulated with the DCFTP-SSA (results not shown). Bimolecular and two-substrate enzymatic reactions are not included in this group and extending the DCFTP or CFTP to provide perfect sampling for the generic SSA of second-order reactions will entail further research.

The DCFTP-SSA introduced here can be viewed as a reformulation and extension of the Gillespie algorithm that provides guaranteed sampling from stationarity for a generic class of systems relevant in GRNs, thereby removing a source of uncontrolled uncertainty from the simulations. By eliminating these extraneous sources of error, comparisons between the predictions of different stochastic models at stationarity can be performed more meaningfully. In the case of the repressilator, we provided an accurate numerical characterization of the stationary distribution for the period of this stochastic oscillator. Because stationarity is guaranteed, the characteristics of this distribution can be correlated with changes in the parameter values (as shown in Fig. 5d) or compared with those of other genetic oscillators. This could be a useful tool for the analysis and design of synthetic transcriptional oscillators and will be the subject of future research.
APPENDIX A: TIME-DEPENDENT SOLUTION OF THE MASTER EQUATION FOR THE ONE-DIMENSIONAL FIRST-ORDER REACTION

The CME for the one-dimensional first-order reaction (Eq. 4) has the same form as a homogenous birth-death process on the nonnegative integers. It is also identical to the equation one obtains when studying the length of a $M/M/\infty$-queue with Poisson arrival and exponential service time (37). Associated with Eq. 4, there exists a family of orthogonal polynomials $(\varphi_j(x))_{j=0}^{\infty}$ with the three-term recurrence relation (52):

$$-x\varphi_j(x) = j\varphi_{j-1}(x) - (k + j)\varphi_j(x) + k\varphi_{j+1}(x),$$

(17)

where $\varphi_0(x) = 1$ and $\varphi_j(x) = 0$ for $j < 0$. Karlin and McGregor showed that the recurrence relation (Eq. 17) leads to a spectral representation for $P_j(t)$, the transition probability of going from state $i$ to state $j$ in time $t$.

$$P_j(t) = P^*_j \int_0^\infty e^{-xt}\varphi_j(x)\varphi_i(x)dx,$$

(18)

where $\varphi(x)$ is a positive measure on $x$ and $P^*_j$ is the stationary distribution.

The recurrence relation (Eq. 17) is satisfied by the Charlier polynomials with $(\varphi(x)$ equal to the Poisson distribution (52), i.e., the Charlier polynomials are a family of polynomials which are orthogonal with respect to the Poisson measure on a discrete lattice on the nonnegative integers (53). The Charlier polynomial of order $i$, parameter $a$, and argument $x$ is denoted $C_i(a,x)$.

Using the following bilinear generating form of the Charlier polynomials (54)

$$\sum_{i=0}^\infty C_i(m;x)C_i(n;y) = e^y(1-\frac{x}{y})^m(1-\frac{z-\frac{y}{z}}{y})^n,$$

(19)

Lee (55) showed that Eq. 18 can be simplified to give

$$P_j(t) = \frac{\exp(-k(1-e^{-t}))}{j!}[k(1-e^{-t})]^j(1-e^{-t})^i \times C_j\left(j;\frac{k(1-e^{-t})^2}{e^t}\right).$$

(20)

The important special case with initial condition given by a $\delta$-distribution at the origin can be simplified even further. If $P_j(0) = \delta(0)$, then $C_{\delta(j,a)} = 1$ and Eq. 20 becomes

$$P_j(t) = \frac{\exp(-k(1-e^{-t}))}{j!}[k(1-e^{-t})]^j = P^*_j(t),$$

(21)

which is a Poisson distribution with time-varying parameter $k(t) = k(1-e^{-t})$. Therefore, the stationary Poisson distribution $P^*_j = k^j e^{-k^j}/j!$ is approached exponentially fast.

Note that

$$\lim_{t \to \infty} P_j(t) = \frac{e^{-k^j}}{j!},$$

(22)

which is another way of showing that the stationary distribution process is Poisson, no matter what the initial condition is.

APPENDIX B: STATIONARY SOLUTION OF THE HILL AND MONOD CMES WITH $\alpha = 1$

The stationary solution for the Hill CME (Eq. 10) with cooperativity factor $\alpha = 1$, which is equivalent to the Michaelis-Menten reaction, can be obtained analytically. This reaction can be viewed as a nonlinear birth-death process, which allows us to write the stationary distribution $P^*_j$ as (37)

$$P^*_j = \frac{\lambda_0\lambda_1 \cdots \lambda_{j-1}}{\mu_1\mu_2 \cdots \mu_j},$$

(23)

where $\lambda_i$ and $\mu_i$ are the birth and death rates of state $i$ and $c$ is a normalization constant. Inserting the values from the CME (Eq. 10), we obtain

$$P^*_j = \frac{k^j \theta \cdots \theta + j - 1}{j!} = \frac{k^{j-1}(\theta + 1)}{j!},$$

(24)

with normalization constant $c = \theta(2\sqrt{k})/\sqrt{k^{\theta/2}}$, where $I_\theta(x)$ are Bessel functions. For the special case of $\theta = 1$, we use a recurrence relation for Bessel functions to simplify the above result to

$$e_0^j = \theta(2\sqrt{k}),$$

Consider now the Monod CME (12) with $\alpha = 1$:

$$P_j = \{ (\theta - 1)I_{\theta+j}P_j + (\theta - 1)I_jP_j \text{ if } j > 1 \}
\begin{cases} \frac{k^j}{j!} & \text{if } j \leq 1 \\ \frac{k^j}{j!} & \text{if } j > 1 \end{cases},$$

(25)

Note that the birth rate of the state $j = 0$ is modified to avoid it becoming absorbing. Using the same strategy as for the Michaelis-Menten CME above, one can find the stationary distribution

$$P^*_j = \frac{k^j \theta \cdots \theta + j - 1}{j!} = \frac{k^{j-1}(\theta + 1)}{j!},$$

(25)

with normalization constant $c = k(1+F_1(2,1+\theta,k))$, where $F_1(2,1+\theta,k)$ is Gauss’s hypergeometric function.

We thank Michael Elowitz for helpful comments about the represillator, and Matthieu Bultelle, Matias Rantalainen, Vincent Rouilly, and Sophia Yaliraki for valuable discussions and feedback. We are also grateful for the constructive comments made by one of the anonymous referees. Research funded by the Engineering and Physical Sciences Research Council of the UK.

REFERENCES


Biophysical Journal 93(2) 401–410