Technical Report CoSBi 14/2008

An Efficient and Exact Stochastic Simulation Method to Analyze Rare Events in Biochemical Systems

Hiroyuki Kuwahara
The Microsoft Research - University of Trento
Centre for Computational and Systems Biology
kuwahara@cosbi.eu

Ivan Mura
The Microsoft Research - University of Trento
Centre for Computational and Systems Biology
mura@cosbi.eu

This is the preliminary version of a paper that will appear in
Journal of Chemical Physics
available at http://jcp.aip.org/jcp/top.jsp

An Efficient and Exact Stochastic Simulation Method to Analyze Rare Events in Biochemical Systems

Hiroyuki Kuwahara, Ivan Mura
Microsoft Research – University of Trento
Centre for Systems and Computational Biology
Trento, 38100, Italy
kuwahara@cosbi.eu, mura@cosbi.eu

July 8, 2008

Abstract

In robust biological systems, wide deviations from highly controlled normal behavior may be rare, yet they may result in catastrophic complications. While *in silico* analysis has gained an appreciation as a tool to offer insights into systems-level properties of biological systems, analysis of such rare events provides a particularly challenging computational problem. This paper proposes an efficient stochastic simulation method to analyze rare events in biochemical systems. Our new approach can substantially increase the frequency of the rare events of interest by appropriately manipulating the underlying probability measure of the system, allowing high-precision results to be obtained with substantially fewer simulation runs than the conventional direct Monte Carlo simulation. Here, we show the algorithm of our new approach, and we apply it to the analysis of rare deviant transitions of two systems, resulting in several orders of magnitude speedup in generating high-precision estimates compared with the conventional Monte Carlo simulation.

1 Introduction

While rare events are, by definition, ones that occur with extremely small probability, they can have significant influences and profound consequences in many systems [2]. This is particularly true in biochemical and physiological systems in that, while the occurrence of biochemical events that leads to some abnormal states may be rare, it may have devastating effects. For example, it has been shown that rare epigenetic modifications play crucial roles in the development of cancer cells by, among other things, inactivating

tumor-suppressing genes [3, 15, 5, 4]. The failed recognition of such dangerous cells by the immune system and the inability to induce apopotosis as a self-defense mechanism are another infrequent yet devastating event, potentially leading to growth and spread of tumors [20]. Thus, gaining insights into the underlying biochemistry of such rare events is crucial for better understanding of the development and physiology of disease.

Since computational methods come with virtually unlimited controllabilities and observabilities of biochemical systems, *in silico* analysis may provide a tool to shed some light on the physiology of such rare yet catastrophic events. The most exact way to analyze a quantitative model of a biochemical system is *molecular dynamics*, where movements of every molecule are tracked [10, 11]. The system state of molecular dynamics consists of the positions and the velocities of every molecule where the dynamics is described by capturing every movement and every collision of molecules. While this approach can describe the time evolution as well as the spatial distribution of each molecule, acquiring such detailed knowledge and performing such computationally expensive simulations is typically infeasible.

Stochastic chemical kinetics (SCK) describes the time evolution of well-stirred biochemical systems as a discrete-space stochastic process. By making the well-stirred assumption, the spatial property of a system can be abstracted away, overriding the system state to be simply the populations of species in the system. This greatly simplifies the complexity of the system state description. The time evolution of the probability distribution of a SCK model is governed by the *chemical master equation* (CME) [19, 9]. However, directly obtaining the solution of the CME of any realistic system, either analytically or numerically, is not feasible owning to its intrinsic complexity. Thus, exact numerical realizations of a SCK model via the *stochastic simulation algorithm* (SSA) [7, 8] are often used to infer the temporal system behavior with a much smaller memory footprint.

Unfortunately, the computational requirements of the SSA can be substantial due to the fact that it requires a potentially large number of simulation runs in order to estimate the system behavior at a reasonable degree of statistical confidence. And, this problem becomes further pronounced in the analysis of rare events as it necessitates generation of a substantial number of sample trajectories. For example, the spontaneous, epigenetic switching rate from the lysogenic state to the lytic state in phage λ-infected *Escherichia coli* [22] is experimentally estimated to be in the order of 10^{-7} per cell per generation [17]. Thus, the SSA would expect to generate sample trajectories of this rare event only once every 10^7 runs, and it would require more than 10^{11} simulation runs to generate an estimated probability with the 95% confidence interval with 1% relative half-width. Therefore, even if the detailed molecular-reaction description of a physiological system were available, a quantitative rare event analysis of such a system might be unfeasible—even with a cluster of thousands of computers.

This paper introduces a new Monte Carlo simulation method to efficiently analyze rare events of biochemical systems. This approach, which we call *weighted SSA* (wSSA), increases the chance to observe the rare events of interest by utilizing the *importance sampling* technique [12]. Importance sampling manipulates the probability distribution of the sampling so that the events of interest can be observed more frequently than it would with the conventional Monte Carlo sampling. The outcome of each biased sampling is weighted by a likelihood factor to yield the statistically correct and unbiased results. Thus, the importance sampling approach can increase the fractions of samples that result in the events of interest per a given set of simulation runs, and consequently, it can efficiently increase the precision of the estimated probability. By applying importance sampling to simulation of biochemical systems, hence, the wSSA can substantially increase the frequency of observation of the rare events of interest, allowing reasonable results to be obtained with orders of magnitude smaller simulation runs than the SSA. This can result in a substantial increase in computational efficiency of rare event analysis of biochemical systems.

The rest of this paper is organized as follows. Section 2 overviews the SCK and its analysis methods including the SSA. Section 3 describes the wSSA while Section 4 discusses the choice of the biased probability measure for the wSSA. The algorithm of the wSSA to analyze rare events is discussed in Section 5. Section 6 presents case studies to compare the accuracy and efficiency of the SSA and the wSSA. Finally, this paper concludes in Section 7 by summarizing the wSSA and describing directions for potential future works.

2 Stochastic Chemical Kinetics

An SCK model is composed of N chemical species $\{S_1, \ldots, S_N\}$ which interact through M *irreversible* reactions $\{R_1, \ldots, R_M\}$ inside a well-stirred, chemically reacting system with a constant volume in thermal equilibrium at some constant temperature. By letting $\mathbf{X}(t) \equiv (X_1(t), \ldots, X_N(t))$ be the system state vector that represents the population of each S_i, SCK describes the time evolution of $\mathbf{X}(t)$ as a discrete-space Markovian process.

In SCK, the occurrence of each reaction R_j is viewed as a discrete random event that changes the system state by $\mathbf{v_j} \equiv (v_{1j}, \ldots, v_{Nj})$, called the *state change vector*, whose i^{th} element v_{ij} specifies the change in X_i by one R_j reaction event. Thus, given the system is in state $\mathbf{x} \equiv (x_1, \ldots x_N)$, the system jumps to state $\mathbf{x} + \mathbf{v_j}$ as a consequence of a single R_j reaction event. The time that the next event of reaction R_j occurs is governed by function a_j, which is called the *propensity function* of reaction R_j, where $a_j(\mathbf{x})dt$ is defined as the probability that, given $\mathbf{X}(t) = \mathbf{x}$, reaction R_j will occur in the next infinitesimal time interval $[t, t + dt)$. With these definitions, SCK

describes the time evolution of $P(\mathbf{x}, t \mid \mathbf{x_0}, t_0)$, the probability of $\mathbf{X}(t) = \mathbf{x}$ given $\mathbf{X}(0) = \mathbf{x_0}$ as:

$$\begin{aligned} P(\mathbf{x}, t + dt \mid \mathbf{x_0}, t_0) = P(\mathbf{x}, t \mid \mathbf{x_0}, t_0) \left[1 - \sum_{j=1}^{M} a_j(\mathbf{x}) dt\right] \\ + \sum_{j=1}^{M} [P(\mathbf{x} - \mathbf{v_j}, t \mid \mathbf{x_0}, t_0) a_j(\mathbf{x} - \mathbf{v_j}) dt]. \end{aligned} \tag{1}$$

Taking the limit for $dt \to 0^+$ and with some algebraic manipulations Equation 1 can be rewritten as the following difference-differential equation:

$$\frac{\partial P(\mathbf{x}, t \mid \mathbf{x_0}, t_0)}{\partial t} = \sum_{j=1}^{M} [P(\mathbf{x} - \mathbf{v_j}, t \mid \mathbf{x_0}, t_0) a_j(\mathbf{x} - \mathbf{v_j}) - P(\mathbf{x}, t \mid \mathbf{x_0}, t_0) a_j(\mathbf{x})], \tag{2}$$

which is called the *chemical master equation* (CME) [19, 9, 24]. Although the time integral of the CME gives the probability $P(\mathbf{x}, t \mid \mathbf{x_0}, t_0)$ for any $t > t_0$, directly obtaining the solution of the CME of nontrivial systems, either analytically or numerically, is not feasible [8, 24, 6]. Thus, owing to its intrinsic complexity, the CME itself may not be useful for analyzing the temporal behavior of biochemical systems without reducing the system state space with approximations such as those described in [16, 21].

In order to more practically analyze the time evolution of $\mathbf{X}(t)$ within an SCK model, a Monte Carlo simulation algorithm called the *stochastic simulation algorithm* (SSA) has been developed [7, 8]. SSA is derived by defining a probability density function $p(\tau, j \mid \mathbf{x}, t)$ such that $p(\tau, j \mid \mathbf{x}, t)d\tau$ is the probability that, given $\mathbf{X}(t) = \mathbf{x}$, the next reaction occurs in the infinitesimal time interval $[t + \tau, t + \tau + d\tau)$, and it is R_j. Then, it can be shown that:

$$p(\tau, j \mid \mathbf{x}, t) = a_0(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau) \times \frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}, \tag{3}$$

where:

$$a_0(\mathbf{x}) \equiv \sum_{j=1}^{M} a_j(\mathbf{x}). \tag{4}$$

Thus, formally, the SSA is a Monte Carlo simulation procedure that faithfully selects j and τ according to the probability distribution defined in Equation 3. In other words, in the SSA, the time to the next reaction, τ, is a sample of the negative exponential random variable $T(\mathbf{x})$ with mean $1/a_0(\mathbf{x})$ and the index of the next reaction, j, is a sample of the discrete random variable $J(\mathbf{x})$ with probability mass function $p_J(j; \mathbf{x}) = a_j(\mathbf{x})/a_0(\mathbf{x})$, $j = 1, 2, \ldots, M$.

3 Weighted SSA

In the *direct method* implementation of the SSA [7], samples of $J(\mathbf{x})$ are drawn by first picking a unit uniform random value, u, and then choosing the smallest j satisfying:

$$\sum_{\mu=1}^{j} a_\mu(\mathbf{x}) \geq u a_0(\mathbf{x}).$$

This scheme correctly generates independent samples based on $J(\mathbf{x})$. In other words, by letting $[q]$ be the Iverson bracket [13] such that

$$[q] = \begin{cases} 1 & \text{if } q \text{ is true,} \\ 0 & \text{otherwise,} \end{cases}$$

$p_J(j;\mathbf{x})$ can be expressed as:

$$p_J(j;\mathbf{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} [J^{\{i\}}(\mathbf{x}) = j] \tag{5}$$

where $J^{\{1\}}(\mathbf{x}), \cdots, J^{\{n\}}(\mathbf{x})$ is a sequence of n independent samples from the next reaction selection scheme in the direct method, given that $\mathbf{X} = \mathbf{x}$. Equation 5 implicitly shows that each sample is of equal weight, and hence, the j-th reaction event is expected to occur once every $1/p_J(j;\mathbf{x})$ samples. An implication of this is that, if $p_J(j;\mathbf{x}) \approx 0$, it is highly likely that observing even a single j-th reaction requires a very large sample size. Thus, the SSA typically requires a substantial number of simulation runs before state transitions led by a sequence of such rare reactions are observed. This presents significant computational demands for analysis of rare events in biological systems via the SSA.

In order to alleviate the computational requirements in the analysis of rare events of a SCK model, the *weighted SSA* (wSSA) uses the importance sampling technique. In importance sampling, the average of a function of a random variable Y on Ω with density function $p_Y(y)$ is estimated using a different random variable $\bar{Y}$ on Ω with density function $p_{\bar{Y}}(y)$. Thus, let $\langle [Y \in E] \rangle$, the average of $[Y \in E]$, be the property of interest. Then, by definition,

$$\langle [Y \in E] \rangle = \int_{-\infty}^{\infty} [y \in E]\, p_Y(y) dy. \tag{6}$$

Multiplying and dividing the right hand side by $p_{\bar{Y}}(y)$ yields:

$$\begin{aligned} \langle [Y \in E] \rangle &= \int_{-\infty}^{\infty} \frac{[y \in E]\, p_Y(y)}{p_{\bar{Y}}(y)} p_{\bar{Y}}(y) dy \\ &= \langle \frac{\left[\bar{Y} \in E\right] p_Y(\bar{Y})}{p_{\bar{Y}}(\bar{Y})} \rangle. \end{aligned} \tag{7}$$

Since $\langle [Y \in E] \rangle$ is identical to, $P(Y \in E)$, the probability that $Y \in E$, this shows that $P(Y \in E)$ can be expressed using $\bar{Y}$. That is, $P(Y \in E)$ can also be expressed via sampling of $\bar{Y}$ as:

$$P(Y \in E) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\left[\bar{Y}^{\{i\}} \in E\right] p_Y(\bar{Y}^{\{i\}})}{p_{\bar{Y}}(\bar{Y}^{\{i\}})}, \tag{8}$$

where $\bar{Y}^{\{i\}}$ is the i-th independent sample of $\bar{Y}$. Hence, if $\bar{Y}$ has a higher probability for E to occur, then this approach has a higher chance to generate samples $\bar{Y}^{\{i\}}$ such that $\bar{Y}^{\{i\}} \in E$.

Now, consider the next reaction selection scheme of the direct method, and let $P_k(j_k, k; \cdots; j_2, 2; j_1, 1 \mid \mathbf{x_0})$ denote the probability that, given $\mathbf{X} = \mathbf{x_0}$, the first reaction is R_{j_1}, the second reaction is R_{j_2},..., and the k-th reaction is R_{j_k}. Then, since $\mathbf{X}(t)$ is Markovian, this joint conditional probability can be expressed using $p_J(j; \mathbf{x})$ as follows:

$$P_k(j_k, k; \cdots; j_2, 2; j_1, 1 \mid \mathbf{x_0}) = \prod_{h=1}^{k} p_J(j_h; \mathbf{x_{h-1}}) \tag{9}$$

where $\mathbf{x_h} = \mathbf{x_0} + \sum_{h'=1}^{h-1} \mathbf{v_{j_{h'}}}$. Here, let $\bar{J}(\mathbf{x})$ be a biasing discrete random variable with probability distribution $p_{\bar{J}}(j; \mathbf{x})$ in the wSSA. Because $p_J(j; \mathbf{x})$ can be rewritten as $w(j; \mathbf{x}) p_{\bar{J}}(j; \mathbf{x})$ where $w(j; \mathbf{x}) = p_J(j; \mathbf{x}) / p_{\bar{J}}(j; \mathbf{x})$, Equation 9 can also expressed as follows:

$$P_k(j_k, k; \cdots; j_2, 2; j_1, 1 \mid \mathbf{x_0}) = \prod_{h=1}^{k} w(j_h; \mathbf{x_{h-1}}) p_{\bar{J}}(j_h; \mathbf{x_{h-1}}). \tag{10}$$

Consequently, with n runs of Monte Carlo simulation via the wSSA, this k-step path probability can be estimated as:

$$\bar{P}_k(j_k, k; \cdots; j_2, 2; j_1, 1 \mid \mathbf{x_0}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{h=1}^{k} w(j_h, \mathbf{x_{h-1}}) \left[\bar{J}(\mathbf{x_{h-1}}) = j_h\right]. \tag{11}$$

Hence, with an adequate choice of the biasing distribution $p_{\bar{J}}(j; \mathbf{x})$, the wSSA can increase the fractions of sample trajectories that result in the rare events of interest. And each such outcome is weighted by a score

$$\prod_{h=1}^{k} w(j_h, \mathbf{x_{h-1}}),$$

to correct the sampling bias and yield the statistically unbiased results.

4 Predilection Functions

To specify the explicit form of $p_{\bar{J}}(j;\mathbf{x})$ in the wSSA, let $b_j(\mathbf{x})$ denote a function such that $b_j(\mathbf{x})dt$ is the probability with which, given $\mathbf{X} = \mathbf{x}$, one R_j reaction event should occur within the next infinitesimal time dt, based on the bias one might have to lead the system towards the events of interest. Thus, we call $b_j(\mathbf{x})$ the *predilection function* of reaction R_j. With the definition of predilection functions, $p_{\bar{J}}(j;\mathbf{x})$ can be expressed as:

$$p_{\bar{J}}(j;\mathbf{x}) = \frac{b_j(\mathbf{x})}{b_0(\mathbf{x})}, \tag{12}$$

where $b_0(\mathbf{x}) \equiv \sum_{\mu=1}^{M} b_\mu(\mathbf{x})$. Thus, in the wSSA, samples of $\bar{J}(\mathbf{x})$ can be drawn by first picking a unit uniform random value, u, and then choosing the smallest j satisfying:

$$\sum_{\mu=1}^{j} b_\mu(\mathbf{x}) \geq u b_0(\mathbf{x}),$$

while the weight functions $w(j,\mathbf{x})$ can be expressed through the propensity functions and predilection functions as:

$$w(j,\mathbf{x}) = \frac{a_j(\mathbf{x})b_0(\mathbf{x})}{a_0(\mathbf{x})b_j(\mathbf{x})}. \tag{13}$$

In this paper, we have further restricted the form of each predilection function, and defined each predilection function to be proportional to the corresponding propensity function. In other words, for each reaction R_j, $b_j(\mathbf{x})$ is defined as:

$$b_j(\mathbf{x}) = \alpha_j \times a_j(\mathbf{x}), \tag{14}$$

where each $\alpha_j > 0$ is a constant. This allows us to conveniently constrain the predilection functions such that, for each $b_j(\mathbf{x})$, $b_j(\mathbf{x}) = 0$ if and only if $a_j(\mathbf{x}) = 0$. This constraint can avoid the case where a possible trajectory of a system is weighted by a factor 0.

Clearly, if $\alpha_j = \alpha$ for all j, then $p_{\bar{J}}(j;\mathbf{x}) = p_J(j;\mathbf{x})$ and $w(j,\mathbf{x}) = 1$ for all j. Thus, such a selection of predilection functions may not be useful. While optimized selection schemes of the predilection functions require further investigation, it is somewhat intuitive to select predilection functions to alleviate the computational demands in a number of cases. For example, suppose we are interested in analyzing the probability that a species S transitions from θ_1 to θ_2 where $\theta_1 < \theta_2$. Then, most likely, increasing the predilection functions of the production reactions of S and/or decreasing the predilection functions of the degradation reactions of S—even with a small factor—would increase the fractions of the sample trajectories that result in the event of interest.

5 Algorithm

Algorithm 1 implements the wSSA. This algorithm performs n wSSA simulation runs to estimate the probability that the system moves to a state in $\mathcal{E}$—which we presume to be a rare event—within the time limit t_{max}. Note that, while Algorithm 1 is presented in a similar fashion as the counterpart direct method of the SSA, various optimization techniques of the direct method, such as [1, 18], can also be applied to an implementation of the wSSA to further reduce the simulation cost.

First, the algorithm initializes to 0 the variable q, whose value divided by n at the end of the simulation provides the estimate of the probability of interest (line 1). Then, it generates n sample trajectories of $\mathbf{X}(t)$ using the Monte Carlo simulation method of the wSSA. For each simulation run, the initialization is first performed to set the weight of each sample trajectory, w, the time, t, and the system state, $\mathbf{x}$ to 1, 0, and $\mathbf{x_0}$, respectively (lines 3-5). It then evaluates all the propensity functions $a_j(\mathbf{x})$ and all the predilection functions $b_j(\mathbf{x})$, and also calculates $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$ (line 6). Each Monte Carlo simulation is run up to time t_{max}. If a rare event (i.e., $\mathbf{x} \in \mathcal{E}$) occurs within that time frame, then the current sample trajectory weight w is added to q, and the algorithm carries on to the next simulation (lines 8-11). Otherwise, the waiting time to the next reaction, τ, is sampled in the same way as in the direct method of the SSA, and also the next reaction R_μ is selected using the predilection functions (lines 12-14). Then, the algorithm updates the variables, w, t, and $\mathbf{x}$ to reflect the selections of the waiting time and the next reaction (lines 15-17). Any propensity functions and predilection functions which need to be updated based on the firing of one R_μ reaction event are re-evaluated, and $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$ are re-calculated (line 18). After n sample trajectories are generated via the Monte Carlo simulation method, the probability that the system reaches states in $\mathcal{E}$ within t_{max} given the system is in $\mathbf{x_0}$ at time 0 is estimated by q/n (line 21).

The computational complexity of Algorithm 1 and the counterpart of the standard SSA can be compared by noticing that the multiplication/division operations in the wSSA only increases linearly. Indeed, those operation counts in Algorithm 1 differ from the counterpart of the SSA only in the two steps: line 15; and line 18 inside the **while** loop. Line 15 adds a constant number of operations (i.e., 2 multiplications and 2 divisions), while line 18 includes the operations for the update of the predilection functions $b_j(\mathbf{x})$, $j = 1, 2, \ldots, M$ as well as $b_0(\mathbf{x})$. The cost of such updates depends on the specific form of the predilection functions and the network of the model. However, if, as considered in this paper, the predilection functions take the form of simple scaling functions of the propensity functions, then these updates require at most $M + 1$ multiplications, which does not change the overall complexity of the simulation algorithm between the wSSA and the direct method of the SSA.

Algorithm 1 The wSSA

$q \leftarrow 0$
for $k = 1$ to n **do**
  $w \leftarrow 1$
  $t \leftarrow 0$
  $\mathbf{x} \leftarrow \mathbf{x_0}$
  evaluate all $a_j(\mathbf{x})$ and $b_j(\mathbf{x})$, and calculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$
  while $t \leq t_{max}$ **do**
    if $\mathbf{x} \in \mathcal{E}$ **then**
      $q = q + w$
      break out of the while loop
    end if
    $\tau \leftarrow$ a sample of exponential random variable with mean $1/a_0(\mathbf{x})$
    $u \leftarrow$ a sample of unit uniform random variable
    $\mu \leftarrow$ smallest integer satisfying $\sum_{i=1}^{\mu} b_i(\mathbf{x}) \geq u b_0(\mathbf{x})$
    $w \leftarrow w \times (a_\mu(\mathbf{x})/b_\mu(\mathbf{x})) \times (b_0(\mathbf{x})/a_0(\mathbf{x}))$
    $t \leftarrow t + \tau$
    $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{v}_\mu$
    update $a_j(\mathbf{x})$ and $b_j(\mathbf{x})$, and re-calculate $a_0(\mathbf{x})$ and $b_0(\mathbf{x})$
  end while
end for
report q/n as the estimated probability

6 Case Studies

This section analyzes the effectiveness of wSSA for the estimation of rare event probabilities by applying it to two simple biochemical systems and comparing the accuracy and the efficiency with the SSA. The probabilities of interest in these two systems can be numerically obtained by constructing the underlying Markov chain and solving for the corresponding probabilities. Thus, accuracy of the wSSA and the SSA can be quantified by comparing the estimated probabilities with the true probabilities obtained with this scheme.

6.1 Single Species Production-Degradation mechanism

The first model consists of two chemical reactions as follows:

$$\begin{aligned} R_1 &: \; S_1 \xrightarrow{k_1} S_1 + S_2, \\ R_2 &: \; S_2 \xrightarrow{k_2} \varnothing. \end{aligned} \tag{15}$$

This model represents a simple system where species S_1 constitutively synthesizes species S_2 at a rate constant k_1 in reaction R_1 while species S_2 is degraded with rate constant k_2 in reaction R_2. The initial state of the system is given by

$$X_1(0) = 1; \text{ and } X_2(0) = 40,$$

and the rate constants are given by:

$$k_1 = 1.0; \text{ and } k_2 = 0.025.$$

In this system, we are interested in evaluating $P_{t\leq 100}(X_2 \to \theta \mid \mathbf{x_0})$, the probability that the system moves from the initial state $\mathbf{x_0}$ to states where $X_2 = \theta$ within time limit 100. Initially, reactions R_1 and R_2 are in equilibrium because $k_1 \times X_1(0) = k_2 \times X_2(0)$. As X_2 increases, $a_2(\mathbf{x})$ increases while $a_1(\mathbf{x})$ stays the same, resulting in a higher probability of R_2 firing to push back X_2. Symmetrically, if X_2 decreases, $a_2(\mathbf{x})$ decreases while $a_1(\mathbf{x})$ stays the same, resulting in a higher probability of R_1 firing to increase X_2. Hence, X_2 in this system tends to stay around 40. Indeed, the stationary distribution of X_2 follows a Poisson distribution with parameter $\lambda = k_1/k_2 = 40$. Thus, the mean of $\lim_{t\to\infty} X_2(t)$ becomes λ. Consequently, if θ is substantially higher than λ, then $P_{t\leq 100}(X_2 \to \theta \mid \mathbf{x_0})$ may become so small that analysis via the SSA becomes unwieldy.

We have applied both the SSA and the wSSA to estimate $P_{t\leq 100}(X_2 \to \theta \mid \mathbf{x_0})$ for four different values of θ: 65; 70; 75; and 80. For each θ, we have estimated the probability with the two methods at each 10^i-th simulation run, $i \in [1, 7]$, and analyzed the changes of the estimate over simulation

runs by comparing each to the "true" probability, which is obtained through numerical solution (see Appendix for detail on the method). In this analysis, to increase the fractions of the events resulting in X_2 reaching θ within 100 time units with the wSSA, the predilection functions are defined as follows:

$$b_1(\mathbf{x}) = \delta a_1(\mathbf{x}),$$
$$b_2(\mathbf{x}) = \frac{1}{\delta} a_2(\mathbf{x}),$$

where $\delta = 1.2$. The results of this analysis are shown in Figures 1 and 2. Figure 1 compares the changes of the estimate of $P_{t\leq 100}(X_2 \rightarrow \theta \mid \mathbf{x_0})$ via the SSA and the wSSA over simulation runs, and quantifies the accuracy by also comparing the estimates to the true probability of $P_{t\leq 100}(X_2 \rightarrow \theta \mid \mathbf{x_0})$. From Figure 1, it is clear that the wSSA performs better than the SSA in terms of accuracy for given simulation runs. Also, the estimate produced via the wSSA can converge to the true probability more rapidly that the one obtained via the SSA. And, as the value of θ increases, the difference in accuracy becomes more pronounced. For example, while the wSSA produces a fine estimate for $P_{t\leq 100}(X_2 \rightarrow 80 \mid \mathbf{x_0})$ by generating 57,444 out of 10^7 simulation runs resulting in $X_2 = 80$, none of the 10^7 simulation runs via the SSA results in $X_2 = 80$ within 100 time units (Figure 1(d)). To measure the convergence rates more precisely, Figure 2 shows the changes of the relative distance of the estimated probability from the exact one with respect to a number of simulation runs for each value of θ. This shows that, while the relative distance is overall a decreasing function of a number of simulation runs, the wSSA converges more rapidly, and the SSA requires more than two orders of magnitude larger simulation runs in order to achieve the same level of accuracy as the wSSA.

We have measured the efficiency of the wSSA by considering the ratio of the computation time between the wSSA and the SSA. We expect this value to be bounded by a constant as the computational complexity of the wSSA only increases linearly compared with the direct method of the SSA, and as shown in Figure 3(a), the ratio appears to be bounded by 1.5. That is, the SSA is at most 1.5 times faster than the wSSA per simulation run. However, provided that the wSSA can achieve the same level of accuracy with orders of magnitude smaller simulation runs compared with the SSA, the wSSA is substantially more efficient than the SSA. For example, whereas the SSA requires 440 seconds of computation time (i.e., 10^7 simulation runs) to estimate $P_{t\leq 100}(X_2 \rightarrow 65 \mid \mathbf{x_0})$ with relative distance of 3.9×10^{-3} to the true value, the wSSA only requires 5.7 seconds of computational time (i.e., 10^5 simulation runs) to estimate this probability with a relative distance of 2.5×10^{-3} (Figure 1(a)).

To better characterize the computational gain obtained with wSSA, we have evaluated the number of runs required by SSA to achieve a given precision ϵ of the estimate. We define the accuracy ϵ as 1 minus the relative

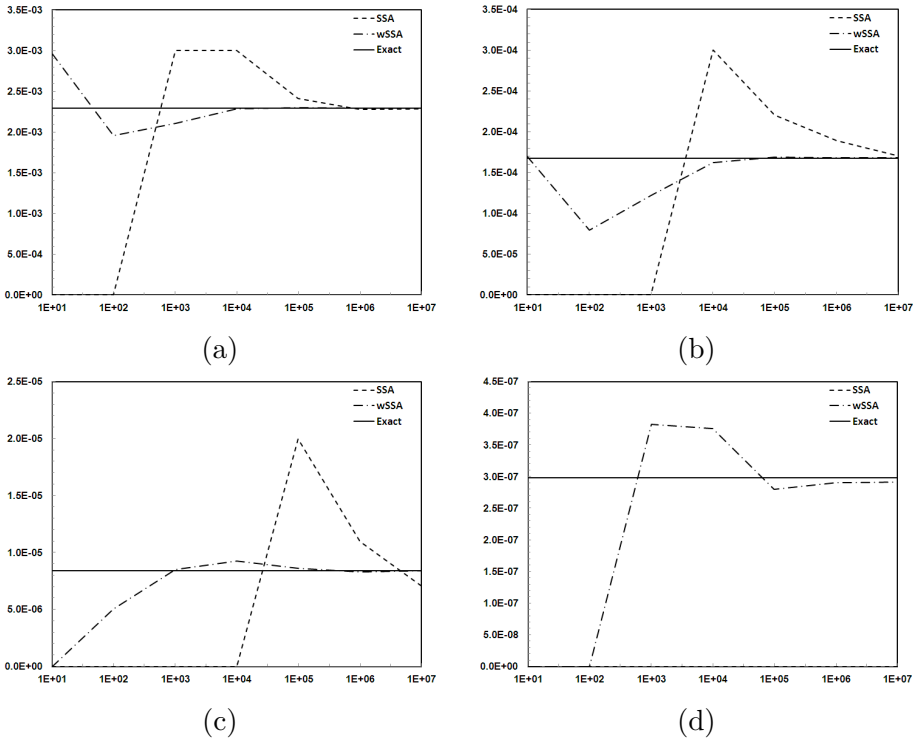


Figure 1: Changes of the estimated $P_{t\leq 100}(X_2 \rightarrow \theta \mid \mathbf{x_0})$ via the SSA and the wSSA with respect to a number of simulation runs for each θ. The solid line represents the true probability. (a) $\theta = 65$, (b) $\theta = 70$, (c) $\theta = 75$, and (d) $\theta = 80$.

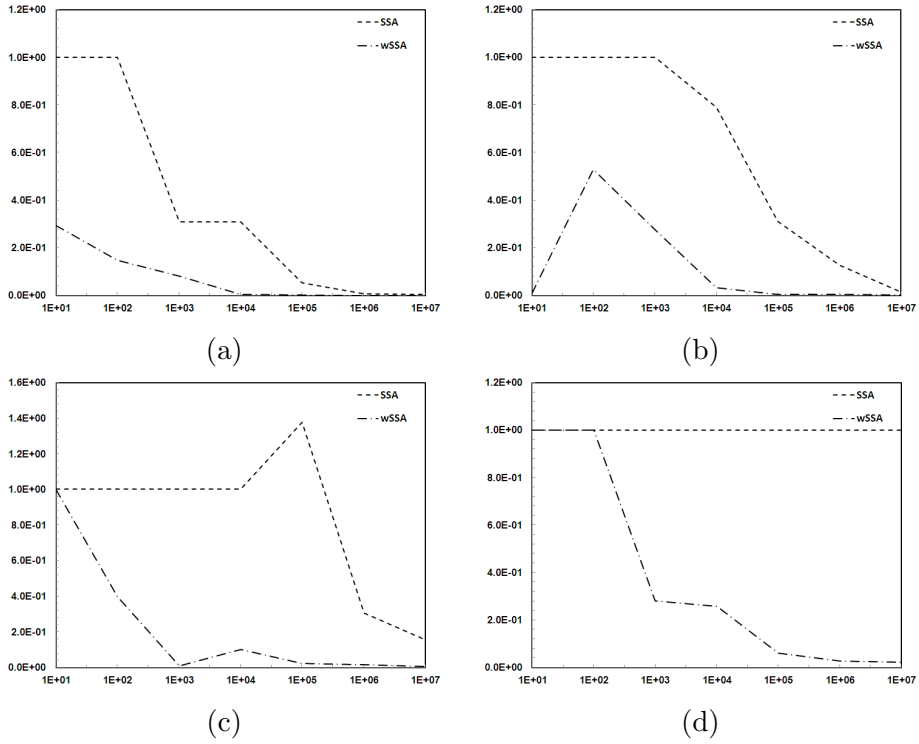
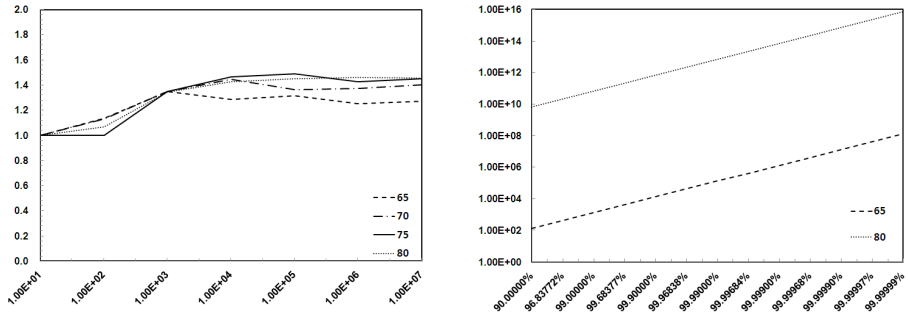


Figure 2: Changes of the relative distance between the true value and the estimate value of $P_{t\leq 100}(X_2 \rightarrow \theta \mid \mathbf{x_0})$ via the SSA and the wSSA with respect to a number of simulation runs for each θ. (a) $\theta = 65$, (b) $\theta = 70$, (c) $\theta = 75$, and (d) $\theta = 80$.



(a) (b)

Figure 3: Comparison between SSA and wSSA computation times. (a) Ratio of the simulation time of wSSA and SSA with respect to a number of simulation runs for the four values of θ in the model given by reactions 15. (b) Ratio of SSA and wSSA computation time (on the vertical axis) for a given level of accuracy ϵ (on the horizontal axis).

distance, and we estimate the number of runs required by SSA through a statistical argument based on confidence intervals (see Appendix for details). Using the estimated number of runs as well as the average run time, we have evaluated the expected computation time of SSA for achieving the accuracy ϵ. Figure 3(b) shows the ratio of the expected computation time between the SSA and wSSA. This clearly indicates that a significant computational gain is achieved with the wSSA algorithm. For instance, while the wSSA can estimate $P_{t\leq 100}(X_2 \rightarrow 80 \mid \mathbf{x_0})$ with an accuracy of 99.9999% in 5.6×10^4 seconds (i.e., with 10^9 simulation runs), to achieve the same level of accuracy with SSA would require around than 2.3×10^5 years of computation (i.e. 1.67×10^{17} simulation runs).

6.2 Enzymatic Futile Cycle Mechanism

The enzymatic futile cycle motif consists of two instances of the elementary single-substrate enzymatic reaction scheme as follows:

$$\begin{aligned}
R_1 &: \; S_1 + S_2 \xrightarrow{k_1} S_3, \\
R_2 &: \; S_3 \xrightarrow{k_2} S_1 + S_2, \\
R_3 &: \; S_3 \xrightarrow{k_3} S_1 + S_5, \\
R_4 &: \; S_4 + S_5 \xrightarrow{k_4} S_6, \\
R_5 &: \; S_6 \xrightarrow{k_5} S_4 + S_5, \\
R_6 &: \; S_6 \xrightarrow{k_6} S_4 + S_2.
\end{aligned} \tag{16}$$

One enzymatic reaction scheme is to transform S_2 into S_5 catalyzed by S_1, and the other one is to transform S_5 into S_2 catalyzed by S_4. This motif can be ubiquitously seen in biological systems including GTPase cycles, mitogen-activated protein kinase cascades, and glucose mobilization [23]. In our model, the initial state of the system is given by

$$X_1(0) = X_4(0) = 1; X_2(0) = X_5(0) = 50; \text{ and } X_3(0) = X_6(0) = 0,$$

and the rate constants are specified as follows:

$$k_1 = k_2 = k_4 = k_5 = 1; \text{ and } k_3 = k_6 = 0.1.$$

Because of the perfect symmetry in the rate constants as well as in the initial molecule counts of the two enzymatic reaction scheme in this system, we expect the system to stay—with high probability—around states in which X_2 and X_5 are balanced. With this model, we are interested in evaluating $P_{t\leq 100}(X_5 \to \theta \mid \mathbf{x_0})$, the probability that, given $\mathbf{X}(0) = \mathbf{x_0}$, the condition $X_5 = \theta$ is satisfied within 100 time units where θ takes four distinct values: 25; 30; 35; and 40.

Since reaction set 16 defines a closed system where the total molecule count is conserved (i.e., $\sum_{i=1}^{6} X_i(t)$ is a constant for all $t \geq 0$), this system has a finite number of states. With our model, the number of the states is relatively small, making the system amenable to a numerical solution of the transient probability distribution of the underlying Markov process (see Appendix for the method). We can therefore compute the exact value of $P_{t\leq 100}(X_5 \to \theta \mid \mathbf{x_0})$, and hence compare the accuracy of the wSSA and the SSA.

In order to increase the fractions of simulation runs that reach the states of interest with the wSSA, we use the following predilection functions:

$$\begin{aligned}
b_1(\mathbf{x}) &= a_1(\mathbf{x}),\\
b_2(\mathbf{x}) &= a_2(\mathbf{x}),\\
b_3(\mathbf{x}) &= \gamma a_3(\mathbf{x}),\\
b_4(\mathbf{x}) &= a_4(\mathbf{x}),\\
b_5(\mathbf{x}) &= a_5(\mathbf{x}),\\
b_6(\mathbf{x}) &= \frac{1}{\gamma} a_6(\mathbf{x}),
\end{aligned}$$

where $\gamma = 0.5$. This biasing can increase the production reaction rate of S_2 while decreasing the production rate of S_5, resulting in an increase in the frequency of X_5 to move to low count states.

Figure 4 shows the estimates of $P_{t\leq 100}(X_5 \to \theta \mid \mathbf{x_0})$ via the SSA and the wSSA with respect to a number of simulation runs for the four values of θ. When $\theta = 40$, the estimate from the wSSA appears to stay in a

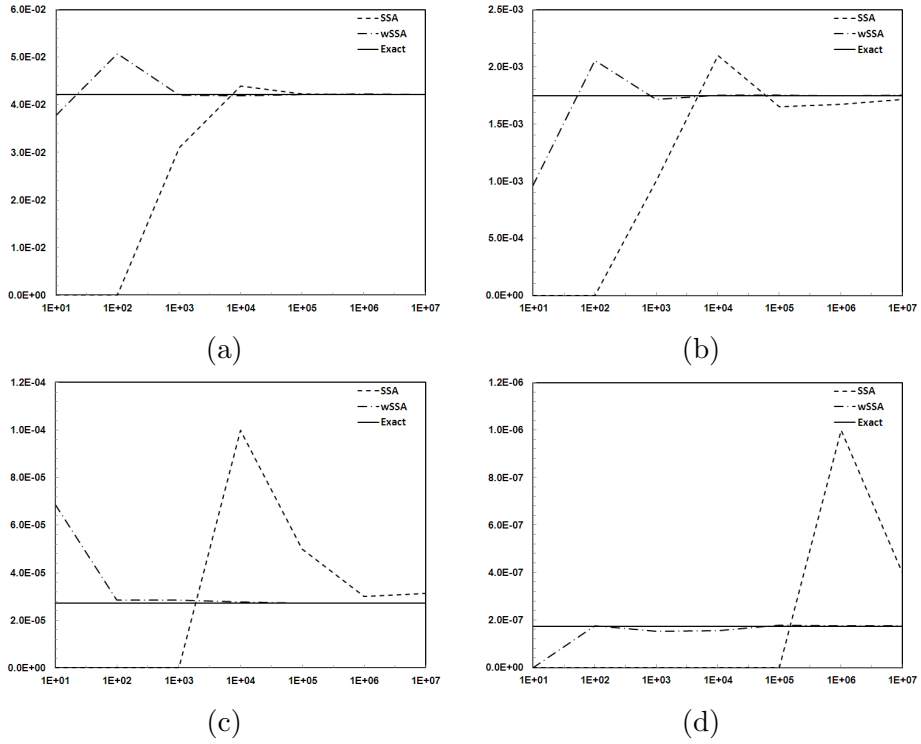


Figure 4: Changes of the estimated $P_{t\leq 100}(X_5 \rightarrow \theta \mid \mathbf{x_0})$ via the SSA and the wSSA with respect to a number of simulation runs for each θ. The solid line represents the true probability. (a) $\theta = 40$, (b) $\theta = 35$, (c) $\theta = 30$, and (d) $\theta = 25$.

good agreement with the true value from as low as 10^3 simulation runs, while the same level of precision requires 10^6 simulation runs with the SSA (Figure 4(a)). This difference becomes more pronounced as the value of θ decreases. For example, when $\theta = 25$, the estimate from the SSA with 10^7 simulation runs cannot achieve the same level of precision that the wSSA obtains with 10^2 simulation runs (Figure 4(d)). To further quantify the convergence rates, Figure 5 shows a comparison of the relative distances of the estimated probability from the exact one with respect to the number of simulation runs. Again, it is shown in the figure that the estimates from the wSSA converge to the true value more rapidly than those from the SSA. Thus, it demonstrates that SSA requires a substantial number of simulation runs to provide a reliable estimate of $P_{t\leq 100}(X_5 \rightarrow \theta \mid \mathbf{x_0})$ compared with the one required by wSSA.

The ratio of the simulation time between the wSSA and the SSA with respect to a number of simulation runs for each θ is illustrated in Figure 6(a). This shows that, in the worst case, the run time of wSSA is 1.2 times

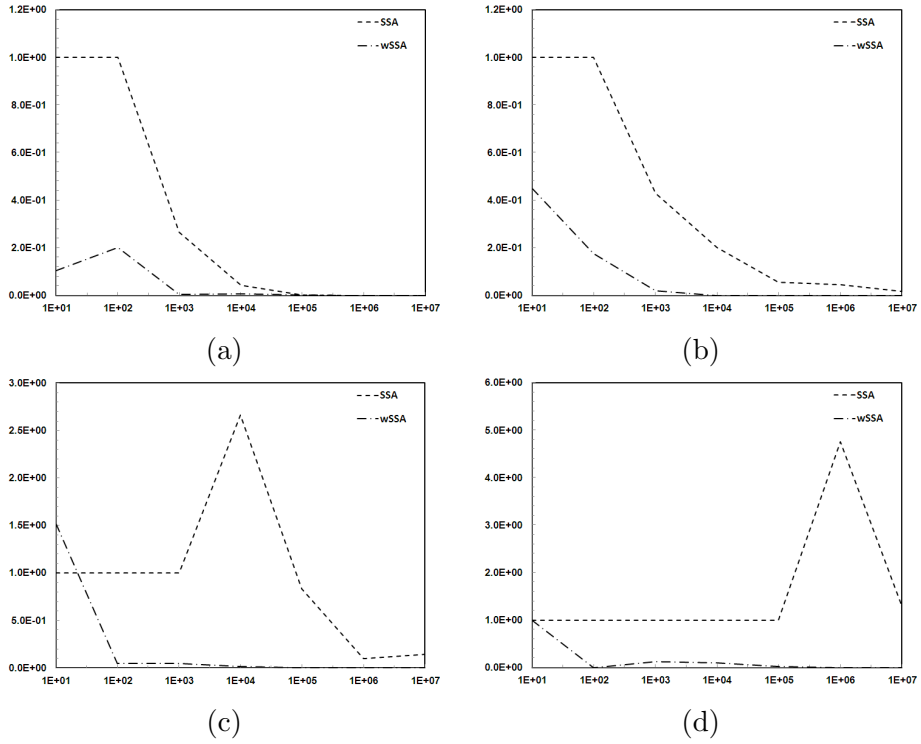
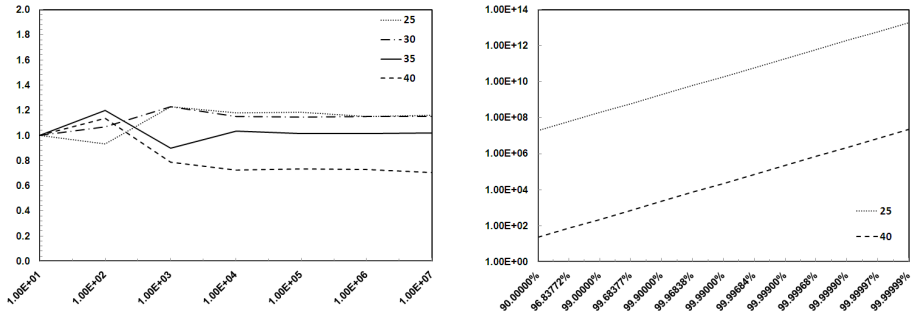


Figure 5: Changes of the relative distance between the true value and the estimate value of $P_{t\leq 100}(X_5 \rightarrow \theta \mid \mathbf{x_0})$ via the SSA and the wSSA with respect to a number of simulation runs for each θ. (a) $\theta = 40$, (b) $\theta = 35$, (c) $\theta = 30$, and (d) $\theta = 25$.



(a) (b)

Figure 6: Comparison between SSA and wSSA computation times. (a) The ratio of the simulation time of the wSSA and the SSA with respect to a number of simulation runs for the four values of θ in the model of Reaction 16. (b) Ratio of SSA and wSSA computation time (on the vertical axis) for a given level of accuracy ϵ (on the horizontal axis).

slower than the direct method of the SSA. In this example, the wSSA is able to get 25% speedup from the counterpart of the SSA when $\theta = 40$. This is because, in the wSSA, higher fractions of sample trajectories result in $X_5 = 40$, allowing many simulation runs to be terminated quickly. Thus, provided the order of magnitude higher precision that the wSSA can achieve per a given number of simulation runs, the wSSA is substantially more efficient in computing $P_{t\leq 100}(X_5 \rightarrow \theta \mid \mathbf{x_0})$ than the SSA. Figure 6(b) shows the ratio of the computation time between with respect to a given level of accuracy of the estimate, which provides further evidence to the higher of wSSA for the estimation of rare event probabilities.

7 Conclusions

This paper has presented a Monte Carlo simulation algorithm (wSSA) to efficiently analyze rare events in biochemical systems by manipulating the underlying probability measure of biochemical systems. This approach facilitates a substantial increase of the fraction of simulation runs that result in the events of interest. Thus, the wSSA can perform high-precision rare event analysis of biochemical and physiological systems with a relatively small number of simulation runs, which would otherwise require a several orders of magnitude larger simulation runs and might take thousands of years of computation with the direct Monte Carlo simulation.

As a case study, we have applied the wSSA to rare event analysis in a simple production-degradation system and a symmetric enzymatic futile cycle system. The preliminary results are promising. In this work, using

a simple biasing scheme of the wSSA, we are able to show that (*i*) the estimate probability from our new simulation approach can rapidly converge to the true probability, and (*ii*) the average run time of a single simulation run via the wSSA only increase linearly with a small constant with respect to the one of the SSA. Therefore, provided that the wSSA can produce a high-precision estimate with orders of magnitude smaller simulation runs compared with the direct Monte Carlo simulation, the wSSA can make the rare event analysis—which would otherwise be infeasible even using a supercomputer—practical on a single PC. Future work includes an optimized selection of predilection functions and more case studies such as analysis of rare deviant effects of physiological systems that can lead a catastrophic complication.

Acknowledgments
The authors would like to thank Daniel T. Gillespie for some helpful discussions. This work was partially supported by the Italian research fund FIRB (project RBPR0523C3).

References

[1] Y. Cao, H. Li, and L. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting system. *Journal of Chemical Physics*, 121:4059–4067, 2004.

[2] M. Csete and J. Doyle. Bow ties, metabolism and disease. *Trends in Biotechnology*, 22(9):446–450, 2004.

[3] G. Egger, G. Liang, A. Aparicio, and P. A. Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463, May 2004.

[4] M. Esteller. Epigenetics in Cancer. *The New England Journal of Medicine*, 358(11):1148–1159, 2008.

[5] A. P. Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143):433–440, May 2007.

[6] C. W. Gardiner. *Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences*. Springer, 3rd edition, 2004.

[7] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[8] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[9] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188:404–425, 1992.

[10] D. T. Gillespie. *Handbook of Materials Modeling*, chapter 5.11, pages 1735–1752. Springer, 2005.

[11] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007.

[12] P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

[13] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1994.

[14] D. Gross and D. R. Miller. The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Operations Research*, 32(2):343–361, 1984.

[15] Y. Jiang, J. Bressler, and A. L. Beaudet. Epigenetics and human disease. *Annual Review of Genomics and Human Genetics*, 5(1):479–510, 2004.

[16] H. Kuwahara, C. Myers, M. Samoilov, N. Barker, and A. Arkin. Automated abstraction methodology for genetic regulatory networks. In *Transactions on Computational Systems Biology VI*, Lecture Notes in Computer Science, pages 150–175. Springer Berlin/Heidelberg, 2006.

[17] J. W. Little, D. P. Shepley, and D. W. Wert. Robustness of a gene regulatory circuit. *EMBO Journal*, 18:4299–4307, 1999.

[18] J. M. McCollum, G. D. Peterson, C. D. Cox, M. L. Simpson, and N. F. Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Computational biology and chemistry*, 30(1):39–49, 2006.

[19] D. A. McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4:413–478, 1967.

[20] J. M. Melief. Cancer: Immune pact with the enemy. *Nature, News and Views*, 450:803–804, 2007.

[21] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *Journal of Chemical Physics*, 124, 2006.

[22] M. Ptashne. *A Genetic Switch*. Cell Press & Blackwell Scientific Publishing, 1992.

[23] M. Samoilov, S. Plyasunov, and A. P. Arkin. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of the National Academy of Sciences US*, 102(7):2310–5, 2005.

[24] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 1992.

[25] P. D. Welch. The statistical analysis of simulation results. In S. Lavenberg, editor, *The Computer Performance Modeling Handbook*, pages 268–328. Academic Press, New York, 1983.

A Computation of Exact Rare Event Probabilities

A.1 Single Species

To compute the probability that the system moves, within the time window $[0, t_{max}]$, from the initial state to a state where the number of molecules is θ, we consider the discrete space Markov process $\{\mathbf{n}\}_t$, whose state represents the number of S_2 molecules at time t, $t \geq 0$. Notice that it is not necessary to represent species S_1 in the state of the model, as its number is constant and equal to 1. The state space of $\{\mathbf{n}\}_t$ is the finite set of integers $\Omega = \{i \mid 0 \leq i \leq \theta\}$, and the transitions rate $q_{i,j}$ from state i to state j, $i, j \in \Omega$, are specified as follows:

$$q_{i,j} = \begin{cases} k_1 & \text{if } j = i+1 \text{ and } i < \theta, \\ jk_2 & \text{if } j = i-1 \text{ and } i < \theta, \\ 0 & \text{otherwise.} \end{cases}$$

The infinitesimal generator matrix of $\{\mathbf{n}\}_t$ is the $(\theta+1) \times (\theta+1)$ matrix Q whose entry i, j is equal to $q_{i,j}$ if $i \neq j$ and to $-\sum_{h \neq i} q_{i,h}$ if $i = j$. The probability distribution vector at time t_{max}, denoted by $\pi(t_{max})$ is given by

$$\pi(t_{max}) = \pi(0) e^{Q t_{max}} \tag{17}$$

where $\pi(0)$ is the initial probability distribution vector, which assigns probability 1 to the state $X_2(0)$ and 0 to all other states. Because state θ is by construction an absorbing one, its component of the mass probability distribution vector $\pi(t_{max})$ is exactly the probability that the process reaches state θ within t_{max}. To compute $\pi(t_{max})$ numerically, we applied the uniformization method [14] for the transient solution of Markov processes.

A.2 Futile Cycle

To compute the probability that the system moves, within the time window $[0, t_{max}]$, from the initial state to a state where the number of S_5 molecules is θ, we consider the discrete space Markov process $\{\mathbf{n}\}_t$, whose state is a 4-dimensional vector $(n_{S_2}, n_{S_3}, n_{S_5}, n_{S_6})$ representing the number of molecules of species S_2, S_3, S_5, S_6 at time t, $t \geq 0$. Notice that the state does not include an explicit representation of the number of S_1 and S_4 molecules as this information is obtained from the value of the components for S_3 and S_6, as $S_1 + S_3 = 1$ and $S_4 + S_6 = 1$. The state space of $\{\mathbf{n}\}_t$ is the following finite set of vectors Ω:

$$\Omega = \left\{ \mathbf{n} \in \mathbb{N}^4 \mid \sum_{i=1}^{4} n_i = X_2(0) + X_5(0) \text{ and } n_{S_2} < \theta, n_{S_3}, n_{S_6} \in \{0,1\} \right\}$$

$$\bigcup \{(\theta, 1, X_2(0) + X_5(0) - \theta, 0), (\theta, 1, X_2(0) + X_5(0) - \theta - 1, 1)\}.$$

The last two states of Ω are absorbing ones, and state transitions from the initial state to those states represent the occurrence of the rare event in the system.

The cardinality of set Ω is $4(X_2(0) + X_5(0) - \theta - 1) + 2$. The transition rates $q_{\mathbf{n},\mathbf{m}}$ from state $\mathbf{n}$ to state $\mathbf{m}$, with $\mathbf{n}, \mathbf{m} \in \Omega$, are as follows:

$$q_{\mathbf{n},\mathbf{m}} = \begin{cases} n_{S_2} k_1 & \text{if } n_{S_3} = 1 - m_{S_3}, m_{S_2} = n_{S_2} - 1, n_{S_2} > \theta, \\ k_{-1} & \text{if } n_{S_3} = 1 - m_{S_3}, m_{S_2} = n_{S_2} + 1, n_{S_2} > \theta, \\ k_2 & \text{if } n_{S_3} = 1 - m_{S_3}, m_{S_5} = n_{S_5} + 1, \\ n_{S_5} k_3 & \text{if } n_{S_6} = 1 - m_{S_6}, m_{S_5} = n_{S_5} - 1, \\ k_{-3} & \text{if } n_{S_6} = 1 - m_{S_6}, m_{S_5} = n_{S_5} + 1, \\ k_4 & \text{if } n_{S_6} = 1 - m_{S_6}, m_{S_2} = n_{S_2} + 1, \\ 0 & \text{otherwise.} \end{cases}$$

The infinitesimal generator matrix Q of $\{\mathbf{n}\}_t$ is defined in the same way as it was defined in Appendix A.1. The initial state probability distribution at time $t = 0$ is the vector $\pi(0)$ that assigns probability 1 to the state $(X_2(0), 0, X_5(0), 0)$ and 0 to all other states. The product $\pi(0)e^{Qt_{max}}$ provides the probability distribution vector of $\{\mathbf{n}\}_t$ at time t_{max}, given the initial state. This computation is again performed through the uniformization algorithmic technique [14]. Because the two states representing the occurrence of the rare event are absorbing ones, the sum of their state probabilities at time t_{max} gives the probability of the system reaching the states of interest.

B Estimation of the Number of SSA Runs

A simple statistical argument can be used to evaluate the expected number of simulation runs via the SSA to provide an estimate within a given relative distance from the exact value p. Let us denote by p_n the estimate of the measure of interest obtained with n simulation runs of SSA. The absolute distance between p_n and p is bounded, with 95% probability, by the half-width of the confidence interval, as follows [25]:

$$|p - p_n| \leq \frac{u\sigma}{\sqrt{n}},$$

where $u = 1.96$ is the 97.5^{th} percentile of the normal standard distribution, p the exact value of the estimated probability and σ the standard deviation of the discrete Bernoulli distribution that gives probability p to the rare event of interest. For a Bernoulli distribution, the variance is $p(1-p)$, hence the standard deviation is $\sigma = \sqrt{p(1-p)}$. The relative distance of p_n from the exact value p is therefore given by the half-width of the confidence interval divided by p. This means that the relative distance of an estimate obtained

with n runs of SSA simulation, is, with 95% probability, less than or equal to

$$\frac{u\sqrt{p(1-p)}}{p\sqrt{n}}.$$

Therefore, in order to obtain an estimate that is within a relative distance of δ, the expected number of SSA simulation runs is expressed as follows:

$$\left\lceil \left(\frac{u\sqrt{p(1-p)}}{p\delta} \right)^2 \right\rceil.$$