Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels

Michael A. Gibson* and Jehoshua Bruck

California Institute of Technology, Department of Computation and Neural Systems, Mail Code 136-93, Pasadena, CA 91125

Received: October 19, 1999; In Final Form: December 14, 1999

There are two fundamental ways to view coupled systems of chemical equations: as continuous, represented by differential equations whose variables are concentrations, or as discrete, represented by stochastic processes whose variables are numbers of molecules. Although the former is by far more common, systems with very small numbers of molecules are important in some applications (e.g., in small biological cells or in surface processes). In both views, most complicated systems with multiple reaction channels and multiple chemical species cannot be solved analytically. There are *exact* numerical simulation methods to simulate trajectories of discrete, stochastic systems, (methods that are rigorously equivalent to the Master Equation approach) but these do not scale well to systems with many reaction pathways. This paper presents the Next Reaction Method, an exact algorithm to simulate coupled chemical reactions that is also *efficient*: it (a) uses only a single random number per simulation event, and (b) takes time proportional to the *logarithm* of the number of reactions, not to the number of reactions itself. The Next Reaction Method is extended to include time-dependent rate constants and non-Markov processes and is applied to a sample application in biology (the lysis/lysogeny decision circuit of lambda phage). The performance of the Next Reaction Method on this application is compared with one standard method and an optimized version of that standard method.

1. Introduction

The process of creating a mechanistic, predictive model of a system can be broken into two steps: (a) creating a complete description of the chemical, physical, and biological processes involved; and (b) using mathematics to generate predictions. For chemical processes, the first step is accomplished by writing a system of chemical equations and perhaps a description of certain physical processes (e.g., temperature, volume, electric field, diffusion, stirring). In principle, (a) can be accomplished without any thought about or care for (b); one may describe a system completely without reference to the subsequent process of calculation. In fact, such a calculation-independent model is more fundamental than calculation-based models: the same process (e.g., $A + B \rightarrow C$) occurs whether there are 20 or 10^{20} molecules of *A* and *B*. The *mathematical* problem, (b), is very much different if there are 20 or 10^{20} molecules of *A* and *B*.[1] Additional mathematical or computational assumptions may be used to simplify calculations. These assumptions are just simplifying computational assumptions; they have nothing to do with the process, but rather with how one represents that process to allow efficient computation.

1.1. Mathematical Descriptions of Chemical Processes. A coupled system of chemical reactions, of the form:

$$X_1 + X_2 \rightarrow X_3 + ...$$

states that one molecule of substance X_1 reacts with one of substance X_2 to give one molecule of substance X_3, etc. A complicated chemical process can be decomposed into a set of many such reactions. (Although it is possible to write arbitrarily high-order reactions, virtually all real systems can be broken up into *elementary* reactions that have at most two reactants and rarely more than three products.)

One may make certain computational assumptions and proceed through to predictions. For example, one may *assume* that there are sufficiently many molecules that the number of molecules can be *approximated* as a continously varying quantity that varies deterministically over time. In this approach, one writes a coupled system of differential equations for the concentration of each substance in terms of the concentrations of all others:

$$\frac{d[X_1]}{dt} = f_1([X_1], [X_2], [X_3] ...)$$
$$\frac{d[X_2]}{dt} = f_2([X_1], [X_2], [X_3] ...) \quad (1)$$

and so on. Solving these differential equations results in the concentration of each substance as a function of time.

Sometimes, one *assumes* that the process in question is fast compared with the time scale of interest, and can be considered to have reached equilibrium. In that case, one replaces the differential equation (eq 1) with the algebraic equations

$$0 = f_1([X_1], [X_2], [X_3] ...)$$
$$0 = f_2([X_1], [X_2], [X_3] ...)$$

and so on, whose solutions give only equilibrium concentrations not dynamics. The assertion "the system is in equilibrium" is recognized as an *assumption*; it is not true in general. In fact, the very statement of eq 1 was based on another assumption,

* Corresponding author. E-mail: {gibson, bruck}@paradise.caltech.edu. Fax: (626) 568-1522.

10.1021/jp993732q CCC: $19.00
Published on Web 02/15/2000

namely, that the number of molecules can be approximated as a *continously* varying quantity that varies *deterministically* over time. This assumption is so fundamental to the majority of chemical kinetics that it is frequently not even viewed as an assumption but as a rigorous consequence of chemical theory. Although this assumption holds for most systems with which one deals, it does not hold in very small systems (which are very important in biology).

Where does that leave us?

One may assume only that all molecules involved obey the laws of quantum mechanics; for systems consisting of complex molecules interacting in complex ways over long times, this approach is completely intractable. One may assume that quantum effects are small and that molecules obey Newton's laws of motion. For very simple systems, this molecular dynamics approach has some merit, but for systems with complex macromolecules (e.g., biological proteins), long time scales, and interactions of several different molecule types, this approach too becomes intractable.

Finally, one may assume that the solution is well mixed—that nonreactive collisions occur far more often than reactive collisions and, hence, that the fast dynamics of motion can be neglected—and one may represent the system simply by the number of each kind of molecule. This approach leads rigorously to the following statement[2]: the probability that a certain reaction μ will take place in the next instant of time dt is given by $a_\mu dt + o(dt)$, where a_μ is independent of dt, and $o(dt)$ denotes terms that are negligible for small dt. However, a_μ may depend on (a) μ, (b) the current number of molecules of each kind, and (c) the current time. (In particular, a_μ depends on temperature and volume, which may change with time.) The remainder of this paper will assume the stochastic framework.

1.2. Objective. An important problem in the stochastic framework is how to develop numerical methods that allow efficient computation. The remainder of this paper presents an exact stochastic simulation algorithm, which we call the Next Reaction Method, that is more efficient than existing methods[3,4] both in terms of number of operations and number of random numbers used. Section 2 provides the theory of the stochastic framework and details of existing algorithms for exact stochastic simulation. Section 3 introduces the Next Reaction Method and its data structures, and proves this algorithm is correct. Section 4 shows how to extend the Next Reaction Method to time-dependent stochastic processes. Section 5 applies the extended algorithm to the Arkin et al.[5] model of lambda phage.

2. Theoretical and Computational Background

This section introduces the theory of the stochastic framework and summarizes existing methods for exact stochastic simulation; that is, numerical methods for performing calculations in the stochastic framework.

2.1. The Stochastic Framework.[2,6] Consider, for example, the set of reactions

$$\begin{aligned} A + B &\xrightarrow{k_1} C \\ B + C &\xrightarrow{k_2} D \\ D + E &\xrightarrow{k_3} E + F \\ F &\xrightarrow{k_4} D + G \\ E + G &\xrightarrow{k_5} A \end{aligned} \qquad (2)$$

The propensities of the reactions are given by k_1, k_2, ... k_5. For example, the probability that a given molecule of A reacts with a given molecule of B in a small time dt is $k_1 dt + o(dt)$. The 'constants' k_i may be a function of volume, temperature, electrolyte concentration, etc.

Remark 1. *Note that the reaction constants given here are* ***not*** *the traditional macroscopic or deterministic rate constants but rather are mesoscopic rate constants, which are related to, but not identical to, macroscopic rate constants.*[3] *In particular, macroscopic rate constants do not depend on volume but concentrations of molecules do, whereas mesoscopic reaction constants do depend on volume but the number of molecules does not.*

One way to proceed is to label each molecule of A (A_1, A_2, ... $A_{\#A}$) and each molecule of B (B_1, B_2, ... $B_{\#B}$), etc. Now there are (#A) × (#B) distinct copies of Reaction 1 that can occur, (#B) × (#C) distinct copies of Reaction 2, etc. (Morton–Firth's algorithm[7] uses this approach; that is, it picks random molecules of A and B and sees whether they react.) Previous work on efficient simulation[4] focused on surface processes, where reactions may take place on a large matrix of (x, y) positions. The current work focuses on reactions in solution, for which position is not important, and one can group *by molecule type*. The (#A) × (#B) copies of Reaction 1 are thus grouped into a single reaction, whose propensity is k_1 × (#A) × (#B) × dt + o(dt). (One can also do such grouping if position is important; but, in that case, grouping by molecule type provides no real benefits because the time-consuming part of the simulation is maintaining the data structures that store position information.)

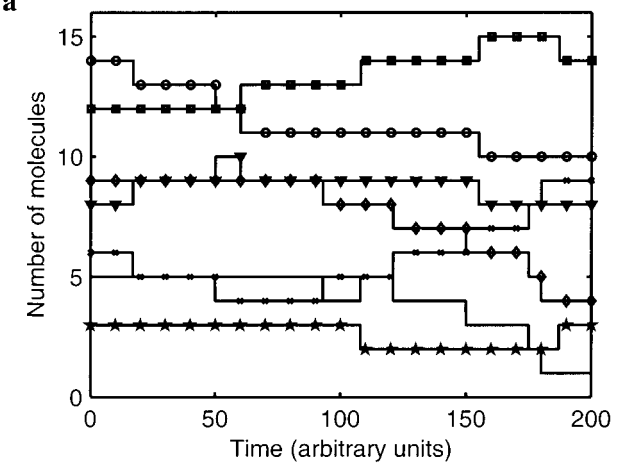
The state of the system in the stochastic framework is defined by the number of molecules of each species and changes discretely whenever one of the reactions is executed. The probability that a certain reaction μ will take place in the next instant of time dt is given by $a_\mu dt + o(dt)$. For example, the state S = (#A, #B, #C, #D, #E, #F, #G) will change to S' = (#A − 1, #B − 1, #C + 1, #D, #E, #F, #G) if Reaction 1 is executed. The probability of this occurrence is given by

$$P(S', t + dt | S, t) = a_1 dt + o(dt) \qquad (3)$$

(Note that because the transition probability depends only on the current state and not on previous states, the underlying process is Markov.) Grouping by molecule type allows immense improvements. Specifically, this paper will demonstrate speed-ups above and beyond those previously reported[4] (again, these are possible because the current paper does not care about positional information). However, there are also cases in which other kinds of groupings are even better; for example, the case study in Section 4.2.1.

One standard way to deal with the stochastic framework is to create one probability variable for each possible state (#A, #B, #C, #D, #E, #F, #G). Then, using eq 3 and the definition of the a_is as a function of state, one writes out a system of coupled differential equations that defines the system. This coupled set of differential equations has *probabilities as variables* and is called a Master Equation. For a system with very few states, such as an ion channel,[8] one may write out this entire system of equations and solve it. For larger systems, however, this approach quickly becomes unreasonable. In the example just given, suppose $0 \le \#A \le 9$, $0 \le \#B \le 9$, etc. Then, there are 10^7 states in this simple 5-equation system. In the Arkin et al.[5] model of lambda phage, reasonable limits on the number of each kind of molecule lead to a number of states

a



b

Time	0	17	50	60	93	108	121	150	155	175
#A	6	5	4	4	5	5	6	7	7	8
#B	14	13	12	11	11	11	11	11	10	10
#C	8	9	10	9	9	9	9	9	8	8
#D	12	12	12	13	13	14	14	14	15	15
#E	9	9	9	9	8	8	7	6	6	5
#F	3	3	3	3	3	2	2	2	2	2
#G	5	5	5	5	4	5	4	3	3	2
Reaction	—	1	1	2	5	4	5	5	2	5

Figure 1. Example trajectory. (a) Graphical representation. Legend: A-X, B-circle, C-triangle, D-square, E-diamond, F-star, G-line. (b) State representation. The 'Reaction' row merely indicates which reaction occured; it is not part of the state.

on the order of 10^{70}. Even if this estimate is off, or a clever reduction of the number of states is possible, say by a factor of 10^{50}, that still leaves 10^{20} states, which is still untractable. Exact stochastic simulation provides a more feasible approach.

2.2. Exact Stochastic Simulation. As a different way of dealing with the stochastic framework, consider the problem of generating a single sample trajectory of a chemical process in the stochastic framework, as in Figure 1a. The (intractable) Master Equation approach tries to write a system of equations and solve simultaneously for the probability of *all* possible trajectories. Generating *a single* trajectory is signficantly easier; as in Figure lb, one needs to generate a sequence of state transitions and the times at which they occur. A naive way to generate legal trajectories is to start with the initial state and repeatedly pick reactions arbitrarily and execute them, thus generating a legal trajectory. A better way to generate trajectories is to pick reactions and times *according to the correct probability distributions* so that the probability of generating a given trajectory with the simulation algorithm is exactly the probability that would come out of the solution of the Master Equation. Amazingly, it is possible to create an algorithm that has this property, *even if* it is not possible to write out the entire Master Equation explicitly, let alone solve it. (There are also *inexact* stochastic simulation algorithms that generate trajectories according to approximately the correct distribution. However, our interest is only in exact methods.)

Given the ability to generate a single trajectory with the correct probability, one may estimate any parameter of interest by generating many trajectories, calculating the value of the parameter for each trajectory, and observing the statistics of those calculated values. For example, to find the average number of molecules of *B* present at time *t*, one can run many trajectories (hundreds or thousands) and plot a histogram of the values of the number of molecules of *B* at time *t*.

Gillespie[3,9] developed two *exact stochastic simulation* algorithms, which are discussed in the next section. The tricky mathematical part of such an algorithm is specifying how to generate random numbers so that they will have the correct distributions. The tremendous success of these exact stochastic simulation algorithms has led to them being applied to much larger systems than was originally anticipated. For example, Arkin et al.[5] used exact stochastic simulation to simulate a model of a simple virus, lambda phage, containing 75 equations in 57 chemical species. Because the original algorithms do not scale readily to large systems, we have developed new versions that do scale well with number of reactions. (The tricky computer science part is to develop *efficient* algorithms that do the right thing.)

2.3. Algorithms. Consider a system of r reactions as in eq 2. For now, assume that all rate constants (e.g., k_1 ... k_5 in eq 2) are true constants; time-varying rate constants will be covered in Section 4. Gillespie[3] proposed two *exact stochastic simulation* algorithms. At each time step, the system is in exactly one state. A transition consists of executing a reaction so there are at most r possible transitions from a given state. The key is to choose random numbers using a computer random number generator and to use those random numbers to pick transitions. One must be careful to choose from the correct distribution at each point in the algorithm.

Gillespie proposed two methods for accomplishing the simulation. The first method, which he calls the *Direct Method*, calculates explicitly *which* reaction occurs next and *when* it occurs. The second method, which he calls the *First Reaction Method*, generates for each reaction μ a putative time τ_μ at which reaction μ occurs, then chooses the reaction μ^* with the smallest τ_μ^* (the *first* reaction) and executes reaction μ^* at time τ_μ^*. Both of these methods are described now.

2.4. Gillespie's Direct Method. For a system in a given state, Gillespie's direct algorithm asks two questions:

- Which reaction occurs next?
- When does it occur?

Clearly, both of these questions must be answered probabilistically by specifying the probability density $P(\mu,\tau)$ that the next reaction is μ and it occurs at time τ. It can be shown[3] that

$$P(\mu,\tau)d\tau = a_\mu \exp(-\tau\sum_j a_j)\, \mathrm{d}\tau \qquad (4)$$

This equation leads directly to the answers of the two aforementioned questions. First, what is the probability distribution for reactions? Integrating $P(\mu,\tau)$ over all τ from 0 to ∞ results in

$$\Pr(\text{Reaction} = \mu) = a_\mu / \sum_j a_j \qquad (5)$$

Second, what is the probability distribution for times? Summing $P(\mu,\tau)$ over all μ results in

$$P(\tau)\mathrm{d}\tau = (\sum_j a_j) \exp(-\tau\sum_j a_j)\, \mathrm{d}\tau \qquad (6)$$

These two distributions lead to Gillespie's direct algorithm:[3]

Algorithm 1. *Exact Stochastic Simulation — Direct Method (Gillespie)*

1. *Initialize(i.e., set initial numbers of molecules, set* $t \leftarrow 0$*).*
2. *Calculate the propensity function,* a_i*, for all i.*
3. *Choose* μ *according to the distribution in eq 5.*

4. Choose τ according to an exponential with parameter $\sum_j a_j$ (as in eq 6).
5. Change the number of molecules to reflect execution of reaction μ. Set $t \leftarrow t + \tau$.
6. Go to Step 2.

As written, this algorithm uses two random numbers per iteration, takes time proportional to the number of reactions to update the a_is, and takes time proportional to the number of reactions to calculate $\sum_j a_j$ and to generate a random number according to the distribution in eq 5. The ideas in the rest of this paper can be used to make the algorithm more efficient so that the time it takes is proportional to the logarithm of the number of reactions (see discussion in the *Appendix*). The rest of the paper will focus ways of improving the First Reaction Method.

2.5. Gillespie's First Reaction Method. The algorithm of the previous subsection is direct in the sense that it generates μ and τ directly. Gillespie also developed the First Reaction Method,[9] which generates a putative time τ_i for each reaction to occur — a time the reaction would occur if no other reaction occurred first — then lets μ be the reaction whose putative time is first, and lets τ be the putative time τ_μ. Formally, the algorithm for the First Reaction Method is as follows:

Algorithm 2. *(Exact Stochastic Simulation — First Reaction Method)*

1. Initialize(i.e., set initial numbers of molecules, set $t \leftarrow 0$).
2. Calculate the propensity function, a_i, for all i.
3. For each i, generate a putative time, τ_i, according to an exponential distribution with parameter a_i.
4. Let μ be the reaction whose putative time, τ_μ, is least.
5. Let τ be τ_μ.
6. Change the number of molecules to reflect execution of reaction μ. Set $t \leftarrow t + \tau$.
7. Go to Step 2.

At first glance, these two algorithms may seem very different, but they are provably equivalent;[9] that is, the probability distributions used to choose μ and τ are the same. We shall not repeat the proof here. As written, this algorithm uses r random numbers per iteration (where r is the number of reactions), takes time proportional to r to update the a_is, and takes time proportional to r to identify the smallest τ_μ.

3. The Next Reaction Method

The following three activities (which occur during every iteration of Gillespie's First Reaction Method) take time proportional to the number of reactions, r: (1) updating all r of the a_is; (2) generating a putative time, τ_i, for each i; and (3) identifying the smallest putative time, τ_μ.

The *Next Reaction Method* will do away with each of these activities in turn. The main ideas used are:

- *Store τ_i, not just a_i.*
- *Be extremely sensitive in recalculating a_i (and τ_i); recalculate a_i only if it changes.* The preceeding statement may seem circular: how can one know that a_i has changed or not changed without calculating it and comparing to its old value? In fact, one can analyze the set of reactions beforehand and determine which reactions change which a_is. Section 3.1 will introduce a data structure, called a *dependency graph*, which allows one to update the minimum number of a_is.
- *Re-use τ_is where appropriate.* In general, Monte Carlo simulations assume statistically independent random numbers, so it is *usually not* legitimate to re-use random numbers. In this particular special case, we shall prove that it *is* legitimate. Specifically, Theorem 1 in Section 3.4 plus two simple transformations make it possible to re-use all τ_is except for τ_μ, the time of the reaction that was just executed. Because re-using random numbers is not valid in general, it is critically important to justify such re-use in this specific case.
- *Switch from relative time (time between reactions) to absolute time.* This switch will obviate the need for one of the two aforementioned transformations: for reactions whose underlying a_i has not changed, the putative time τ_i will not have to change either.
- *Use appropriate data structures to store a_is (and τ_is) so that updating those that change will be a very efficient operation.* Section 3.2 shows a data structure, called an *indexed priority queue*, that achieves this goal.

The formal statement of the algorithm is in Section 3.3, following the definitions of the data structures used.

TABLE 1: Illustration of Definitions 1−3

reaction	a_μ	*DependsOn*(a_μ)	*Affects*(μ)
$A + B \xrightarrow{k_1} C$	$k_1 \times (\#A) \times (\#B)$	A, B	A, B, C
$B + C \xrightarrow{k_2} D$	$k_2 \times (\#B) \times (\#C)$	B, C	B, C, D
$D + E \xrightarrow{k_3} E + F$	$k_3 \times (\#D) \times (\#E)$	D, E	D, F
$F \xrightarrow{k_4} D + G$	$k_3 \times (\#F)$	F	D, F, G
$E + G \xrightarrow{k_5} A$	$k_5 \times (\#E) \times (\#G)$	E, G	A, E, G

3.1. Dependency Graphs. Consider, once again, the reactions in eq 2.

Definition 1. *Let* Reactants(ρ) *and* Products(ρ) *be the sets of reactants and products, respectively, of reaction ρ. So, for example,* Reactants(*Reaction 1*) = {A, B} *and* Products*(Reaction 1)* = {*C*}.

Definition 2. *Let* DependsOn(a_μ) *be the set of substances that affect the value a_μ.*

Evidently, *Reactants(μ)* = *DependsOn(a_μ)*. It is sometimes useful to add additional dependencies (e.g., in the lambda model of Section 5), so we make this distinction.

Definition 3. *Let* Affects(μ) *be the set of substances that change quantity when reaction μ is executed.*

Typically, *Affects(μ)* = *Reactants(μ)*∪*Products(μ)*; but again, there may be exceptions (e.g., catalytic reactions, such as Reaction 3). Table 1 illustrates each of these concepts.

Definition 4. (Dependency Graph). *Let a set of reactions R be given. Let $G(V, E)$ be a directed graph with vertex set $V = R$ and with a directed edge from v_i to v_j if and only if* Affects(v_i) ∩ DependsOn(a_{v_j}) ≠ ∅. *(If for some strange reason, the self edges from v_i to v_i are not included in this definition, include them as well.) Then G is called the dependency graph of the set of reactions R.*

In other words, a dependency graph is a data structure that tells *precisely* which a_is to change when a given reaction is executed. Using the dependency graph allows one to recalculate only the minimum number of a_is in Step 5 of the Next Reaction Method. The dependency graph of the sample reactions is illustrated in Figure 2.

3.2. Indexed Priority Queues. Typically, the dependency graph is *sparse* (i.e., the number of edges from a given vertex is small), and only a few propensities will need to be updated at each time step. It is important to have data structures that are very efficient at handling *a small number* of updates.

The Next Reaction Method deals with two kinds of variables, τ_is and a_is. The latter are easy to handle: the operations required are READ and UPDATE, and they can be stored in a simple array. (A purist might not even store them, but rather recalculate

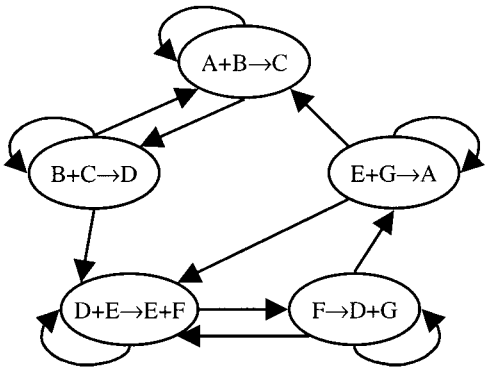


Figure 2. Dependency graphs for example equations from Table 1.

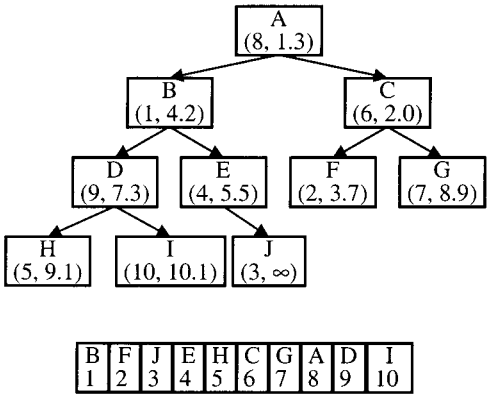


Figure 3. Example indexed priority queue. Top: tree structure. The positions in the tree structure are labeled with letters A−J for pedigogical purposes. Bottom: Index structure. Each number has a pointer to the corresponding position in the tree structure; these pointers are illustrated as letters A−J.

them as needed.) The τ_is require the operations FIND_MINIMUM (in Step 2) and UPDATE (in Step 5d). The former operation is one of the standard operations of a priority queue (which is often implemented as a heap), and with a little thought, the other can be implemented in terms of the standard priority queue algorithms ADD_ELEMENT and DELETE_ELEMENT.[10] However, the standard algorithms, although used in some contexts for this speedup,[4] are not really what is called for in this context. A better UPDATE, which takes into account the structure of the data, requires an indexing scheme and a separate UPDATE algorithm.

Definition 5. *An* indexed priority queue *consists of (a) a tree structure of ordered pairs of the form (i, τ_i), where i is the number of a reaction and τ_i is the putative time when reaction i occurs, and (b) an index structure whose i^{th} element is a pointer to the position in the tree that contains (i, τ_i). The tree structure in (a) has the property that each parent has a lower τ_i than either of its children.*

Figure 3 shows an example of an indexed priority queue. Note the following: (a) finding the minimum element takes constant time — it is always in the top node; (b) the ordering is *only* vertical, not horizontal; (c) the number of nodes is precisely the number of reactions *r*, not twice the number of reactions as in the efficient version of the Direct Method in the Appendix; (d) because of the indexing scheme, it is possible to find any arbitrary reaction in constant time, and (e) $\tau_3 = \infty$, which corresponds to Reaction 3 never occuring (i.e., $a_3 = 0$). In fact, ∞ is a perfectly legitimate floating point number, so it is possible to implement this feature (in the C programming language, for example) without any major headaches.

There are several algorithms that need to be defined to use the indexed priority queue. Most of them are analogous to algorithms for standard priority queues.[10] In particular, one needs

- SWAP (*i, j*), which swaps the tree nodes *i* and *j and updates the index structure appropriately*;
- BUILD, which takes a tree and an index structure and moves entries until the tree has the property that each parent is less than its children; and
- UPDATE (*r*), which updates a given reaction number.

SWAP is easy to implement. BUILD is completely analogous to the standard heap/priority queue BUILD operation but uses SWAP to keep the index structure correct. UPDATE is nonstandard and deserves comment.

Algorithm 3. *UPDATE(node **n**, value **new_value**)*
*Change value of **n** to **new_value***
*UPDATE_AUX(**n**)*

Algorithm 4. *UPDATE_AUX(node **n**)*
*If value(**n**) < value(parent(**n**))*
*SWAP **n** and parent(**n**)*
*Update_aux(parent(**n**))*
*Else If value(**n**) > minimum value(children(**n**))*
*SWAP **n** and minimum child(**n**)*
*Update_aux(minimum child(**n**))*
Else
Return

An example of the algorithms will provide clarification.

Example 1. *Suppose the value of τ_1 changes from 4.2 to 16. Looking in the index array, τ_1 is stored in node B. In the tree structure, one changes the τ value of node B to 16. Calling UPDATE_AUX on node B, one executes the 'Else If' statement and swaps the ordered pairs in nodes B and E and the corresponding indices (1 and 4) in the array. Calling UPDATE_AUX recursively on node E, one notes that the new value of 16 is in the correct position ($5.5 < 16 < \infty$), so the final 'Else' clause is executed and the algorithm stops with ordered pair (4, 5.5) in node B, (1, 16) in node E, index 'E' in position 1 of the array, and index 'B' at position 4. The rest of the structure remains unchanged.*

The converse case, where the new value is less than the old value, is completely analogous.

One way to implement UPDATE is simply to delete the offending node, and insert a new node with the same reaction number but a different time value. This implementation takes something like 2 log *r* operations. Our approach, on the other hand, changes the node in place, then bubbles it up or down the tree structure until the priority property is re-established. This approach evidently takes log *r*, but has the advantage that if there are a small number of reactions that have fast rate constants compared with the others (e.g., there are r' such reactions), then most of the updates will involve those reactions and take log r' time. This savings occurs because once the algorithm reaches a node that is already in the right spot, it does not continue further. For example, if some of the reactions are "disabled" or "not possible" in the given state and have $a = 0$ and $\tau = \infty$, they will not slow down the computation. This effect can be significant; for example, the chemotaxis system of Morton-Firth[7] contains a large number of reactions that will not be "active" at any given time. Because of these inactive reactions, previous work[7] avoided the standard Gillespie algorithm (i.e., the Direct Method), which grows with the number of reactions, and instead developed one that is not exact but scales with number of molecules because the number of

molecules is much less than the number of possible reactions. Note that our algorithm is not only exact but also scales with the *logarithm* of the number of "*active*" reactions. The Application Section (Section 5) gives some performance numbers.

3.3. Statement of Algorithm and Timing Analysis.

Algorithm 5. *(Exact Stochastic Simulation — Next Reaction Method)*

1. *Initialize:*
   (a) *set initial numbers of molecules, set* $t \leftarrow 0$, *generate a dependency graph* G;
   (b) *calculate the propensity function,* a_i, *for all* i;
   (c) *for each* i, *generate a putative time,* τ_i, *according to an exponential distribution with parameter* a_i;
   (d) *store the* τ_i *values in an indexed priority queue* P.
2. *Let* μ *be the reaction whose putative time,* τ_μ, *stored in* P, *is least.*
3. *Let* τ *be* τ_μ.
4. *Change the number of molecules to reflect execution of reaction* μ. *Set* $t \leftarrow \tau$.
5. *For each edge* (μ, α) *in the dependency graph* G,
   (a) *update* a_α;
   (b) *if* $\alpha \neq \mu$, *set* $\tau_\alpha \leftarrow (a_{\alpha,old}/a_{\alpha,new})(\tau_\alpha - t) + t$ *(see note 11);*
   (c) *If* $\alpha = \mu$, *generate a random number,* ρ, *according to an exponential distribution with parameter* a_μ, *and set* $\tau_\alpha \leftarrow \rho + t$;
   (d) *replace the old* τ_α *value in* P *with the new value.*
6. *Go to Step 2.*

Consider the time used by the algorithm. Step 1 of the Next Reaction Method is only executed once; Steps 2−6 are executed once for each simulation event. Steps 3, 4, and 6 do not depend on the number of reactions, r. Step 2 does not either, because of the properties of indexed priority queues. Step 5 is executed once for every edge (μ, α) in G. Suppose there are k such edges, where k is typically much less than r. Step 5a, executed k times, depends on the number of reactants for each (elementary) reaction so it should take no more than three multiplications (as was explained in the *Introduction*). Step 5b, executed $k - 1$ times, requires an addition, a subtraction, a multiplication, and a division. Step 5c, executed 1 time, requires a call to the random number generator, which can be very slow compared with the other operations discussed (a simple test on our system indicates that a single call to the random number generator takes 10 times as long as a division). Step 5d, executed k times, requires *at most* $2 \log r$ operations, although it may effectively take far fewer (see the discussion in Section 3.2). (Throughout this paper, log means logarithm base 2, as per the typical computer science usage.)

The total number of operations per iteration is at most $c_{2,3,4,5a,6} + c_{5b}(k - 1) + c_{5c} + c_{5d}(k)(2 \log r)$, where each c is a machine specific constant. From a computer science perspective, this is $O(\log r)$; that is, for very large r, only the last term will matter. From a more practical perspective, for r of 50 or 100, the other terms, particularly c_{5c}, may not be negligible. Let us be very clear on this point: the Next Reaction Method works even if k is large but will achieve more of a speedup if k is small relative to the number of reactions. (An equivalent way of saying the same thing is 'if the dependency graph is *sparse*.')

Lukkien et al.[4] discuss ways to improve the Direct Method and the First Reaction Method (a more detailed treatment can be found in Segers[12]). Their improved First Reaction Method, which we shall call the Absolute Time First Reaction Method, consists of switching from relative to absolute times and using a standard priority queue. They conclude that for time-invariant processes, the Direct Method is preferable to the Absolute Time First Reaction Method for two reasons that do not apply to the Next Reaction Method. First, in their domain, in which position is important, it is difficult to do the indexing necessary to implement the efficient update algorithm (Algorithm 3 of Section 3.2); specifically, the time-consuming part of their problem is not the priority queue but rather maintaining the data structures that store position-dependent information, which is irrelevant in the present position-independent context. Second, the Absolute Time First Reaction Method generates too many random numbers. Because typical computer pseudo-random number generators cycle with some regularity, using too many random numbers will quickly exhaust the abilities of the generator and should be avoided with extreme prejudice. (Also, from a purely practical standpoint, generating random numbers is relatively slow.)

Amazingly, the Next Reaction Method uses just a single random number per iteration. Clearly, the optimum would be *exactly* one random number per iteration. Our algorithm is slightly suboptimal in that the initialization step will generate an extra r random numbers and at the end of the algorithm r random numbers will be left over. As the number of iterations increases, this initialization effect becomes negligible by comparison. The only new random number generated, τ_μ, corresponds to the reaction that was just executed and is generated in Step 5c. It is clear that reaction μ requires a new random number because the value of the old random number has been used explicitly, thus reducing it to a sure variable. Section 3.4 will show that it is correct to do the other manipulations in Step 5, so as not to regenerate any other random numbers. For this reason we assert that the Next Reaction Method is superior to the Direct Method.

3.4. Re-using τ_is. This section will demonstrate that the Next Reaction Method, with its switch from relative to absolute times and all of the strangeness in Step 5, is equivalent to the First Reaction Method. This demonstration, necessary to show that the algorithm works and why it works, is somewhat more mathematical than the rest of the paper. The reader whose primary interest is implementing the Next Reaction Method may skip ahead with impunity. As mentioned before, it is usually not legitimate to re-use random numbers; this section will prove that it is permissible in this special case. (In what follows, T_i will denote the *random variable* corresponding to the i^{th} reaction, and τ_i, a number, will denote a *sample* from that random variable.)

One of the differences between the First Reaction Method and the Next Reaction Method is that the former uses relative times, whereas the latter uses absolute times. This difference should not be a stumbling block or a source of confusion. Suppose that during the n^{th} iteration of the First Reaction Method, the random variables are denoted R_α, for $1 \leq \alpha \leq$ (number of reactions). Then $R_\alpha = \text{Exp}(a_\alpha)$, and the density of R_α is given by $P_{R_\alpha}(\tau) = \theta(\tau) a_\alpha \exp(-a_\alpha \tau)$. [The Heaviside function, $\theta(\tau)$, is 0 for $\tau < 0$ and 1 for $\tau \geq 0$.] The corresponding absolute time is given by the random variable $T_\alpha = R_\alpha + t_n$, the sum of the relative time and the variable t during the n^{th} iteration. (The n^{th} iteration ends, and the $n + 1^{\text{st}}$ begins when t is updated in Step 5.) What is the density of T_α? By the random variable transformation (RVT) theorem,[13]

$$P_{T_\alpha}(\tau) = \int_{-\infty}^{\infty} P_{R_\alpha}(\tau')\delta(\tau - [\tau' + t_n])\mathrm{d}\tau' = P_{R_\alpha}(\tau - t_n) = \theta(\tau - t_n) a_\alpha \exp(-a_\alpha(\tau - t_n))$$

or, equivalently,

$$\Pr(T_{\alpha,n} > u) = \begin{cases} \exp(-a_{\alpha,n}(u - t_n)) & \text{if } u > t_n \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

Clearly, an absolute-time version of the First Reaction Method with no other changes would be entirely equivalent to the original relative-time version.

Now we turn our attention to the Next Reaction Method. After Step 1, the random variables follow the distribution in eq 7; t_0 was set to 0 in Step 1a. The real core of the Next Reaction Method is that each subsequent iteration maintains eq 7.

At the risk of being overly mathematical, we state the key property that allows the Next Reaction Method as a theorem:

Theorem 1. *Assume that eq 7 holds at the beginning of Step 2. Then, before Step 5 of the n^{th} iteration, for all $i \neq \mu$, τ_i is distributed according to*

$$\Pr(T_i > u) = \begin{cases} \exp(-a_{i,n}(u - t_{n+1})) & \textit{if } u > t_{n+1} \\ 1 & \textit{otherwise} \end{cases} \quad (8)$$

Proof. By assumption, before Step 2 of the n^{th} iteration, τ_i is distributed according to eq 7. Steps 2 and 3 identify the least τ, namely, τ_μ. The act of identification reduces uncertainty. In particular, T_μ becomes the sure variable τ_μ, and all of the other τ_is must be larger than τ_μ. Hence, each of the other T_is is distributed according to $\Pr(T_i > u | T_i > \tau_\mu)$; which by definition is $\Pr((T_i > u) \text{ AND } (T_i > \tau_\mu))/\Pr(T_i > \tau_\mu)$. There are two cases. In Case 1, for $u > \tau_\mu$, the numerator simplifies to $\Pr(T_i > u)$, and the resulting division is $\exp(-a_{i,n}(u - t_n))/\exp(-a_{i,n}(\tau_\mu - t_n)) = \exp(-a_{i,n}(u - \tau_\mu))$. In Case 2, for $u \leq \tau_\mu$, the numerator simplifies to $\Pr(T_i > \tau_\mu)$. In this case, the numerator cancels the denominator, leaving 1. In Step 4, t_{n+1} is set to τ (which was set to τ_μ in Step 3), so the theorem holds.

Showing that eq 7 is maintained is just a matter of collecting the details:

- For those $i \neq \mu$ whose a_i remains constant from the n^{th} to $n + 1^{\text{st}}$ iteration, $a_{i,n+1} = a_{i,n}$, so eq 8 is equivalent to eq 7. There is no need to change these τ_is in Step 5. In fact, reactions whose a_i does not change are not in the dependency graph, so the τ_is are not changed.
- For those $i \neq \mu$ whose a_i *does* change, τ_i is now distributed according to eq 8. Simply plugging into the RVT theorem shows that the random variable T'_i, constructed by $\tau'_i = (a_{i,n}/a_{i,n+1})(\tau_i - t_{n+1}) + t_{n+1}$, is distributed according to eq 7.[14] What is the intuition behind this transformation? By the theorem, τ_i is distributed according to eq 8. Going back to relative times, $\tau_i - t_{n+1}$ is distributed according to $\text{Exp}(a_{i,n})$. It can be shown (by the RVT, for example) that $(a_{i,n}/a_{i,n+1})\text{Exp}(a_{i,n}) = \text{Exp}(a_{i,n+1})$. Returning to absolute times gives the transformation. This transformation is applied to all the appropriate i values in Step 5b.
- Finally, for $i = \mu$, it is necessary to generate a new random number. Note that the theorem only holds for $i \neq \mu$. For $i = \mu$, the variable T_μ was reduced to a sure variable in Step 3, so a new random variable is needed. That new value is supplied in Step 5c.

One key point has been overlooked thus far: the First Reaction Method requires *statistically independent* random numbers. To complete the correctness argument, it must be shown that the manipulations done by the Next Reaction Method do not introduce any statistical dependencies.

At each step in the algorithm $T_i = f_i(R_i)$, each random variable in the Next Reaction Method is a transformed version of the corresponding random variable in the First Reaction Method and there are no cross dependencies. By the RVT theorem,

$$P_{T_1 \cdots T_N}(\tau_1, \cdots \tau_N) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \{\prod_{i=1}^{N} \theta(r_i) \exp(-a_i r_i)\} \prod_{j=1}^{N} \delta(\tau_j - f_j(r_j))\, \mathrm{d}r_1 \cdots \mathrm{d}r_N$$

$$= \prod_{i=1}^{N} \{\int_{-\infty}^{\infty} \theta(r_i) \exp(-a_i r_i) \delta(\tau_i - f_i(r_i))\, \mathrm{d}r_i\}$$

The "product form" of this joint distribution function tells us that because the *original* variables R_i were statistically independent, then the *transformed* variables T_i are as well.

In summary, Step 1 sets up the T_i according to eq 7. Each subsequent iteration maintains that distribution, without introducing any statistical dependencies between the random variables. Thus, the Next Reaction Method is equivalent to the First Reaction Method and, in turn, to the Direct Method and the Master Equation approach.

4. Extension: Time-Dependent and Non-Markov Processes

The two approaches for efficient calculation of trajectories of chemical reactions in the stochastic framework presented thus far assume that the probability of a reaction μ occurring in a little bit of time dt (a) is given by $a_\mu \times \mathrm{d}t$, where a_μ is a constant, and (b) depends only on the current state, not on the previous state or states of the system. It can be shown[2] that many reasonable chemical systems have these properties.

This section will show how to deal with systems in which (a) and (b) do not hold: in particular, it first relaxes assumption (a) by letting a_μ be a function of time (as is necessary to model systems whose rate "constants" change, due to changing temperature, volume, etc.), and second it relaxes assumption (b), showing how to deal with non-Markov processes. Even though *elementary* reactions in the stochastic framework are Markov [i.e., have property (b)], it is sometimes useful to group consecutive steps to form a composite process. The full model of that process is, of course, still Markov, but if one is only interested in a subset of the variables, the resulting mathematical process is not guaranteed to be Markov.

4.1. Time-Dependent Markov Processes. Consider a system in which the probability of a reaction μ occurring in a little bit of time dt is given by $a_\mu \times \mathrm{d}t$, *but a_μ is a function of time*. For now, assume that the transition probabilities do not depend on history, but only on the current state. For example, Reaction 1 in Table 1 has propensity $a_1 = k_1 \times (\#A) \times (\#B)$. The rate constant k_1 is a function of temperature and of volume. In an engineering system, one typically affects the rate constants by heating or cooling the reaction. In a biological system, cell growth changes the volume. Either of these mechanisms, or others, might change rate constants as a function of time, which requires a modification to the algorithms presented.

In place of the simple exponential distribution, the putative times are distributed according to[13]

$$P_\mu(\tau | S, t_n) = a_\mu(S, \tau) \exp(-\int_{t_n}^{\tau} a_\mu(S, t) \mathrm{d}t) \quad (9)$$

Notice two things: first, it is not easy in general to find a closed form solution of eq 9 for arbitrary functions of time a_μ; and second, for nonconstant a_μ, the resulting answer will not be a simple exponential distribution and hence, will not have

the temporal homogeneity property that $\Pr(T_i > u|T_i > \tau_\mu) = \Pr(T_i > u - \tau_\mu)$. As a consequence of the latter, it will not, in general, be possible to let $t_0 = 0$, so absolute times should be used.

4.1.1. How To Do It: Next Reaction Method, Markov Processes. To extend the Next Reaction Method to arbitrary Markov processes, one simply changes Step 3 to generate τ_i according to the new process.[15] This change has the following two advantages over the time-variant version of the Direct Method in the *Appendix*:

- Because one considers each reaction separately, the computation is easier and may be analytically solvable for some processes (e.g., in the example of the next section). The Direct Method, which considers all reactions at once, involves a sum within the integral in eq 9; the exact form is given in the *Appendix*.
- Because the Next Reaction Method stores τ_is and not just a_is, it does not have to recondition on each iteration. Specifically, after executing reaction μ it does not need to regenerate τ_i according to $\Pr(T_i > u|T_i > \tau_\mu)$. The fact that the algorithm chose to execute reaction μ implies $T_i > \tau_\mu$. Therefore, T_i is already distributed according to the correct distribution for all $i \neq \mu$ whose a_i has not changed.

We now show an example of how to generalize the Next Reaction Method for time-varying Markov processes, then consider the problem of re-using random numbers with the generalization.

4.1.2. Example: Changing Volume. Reaction 1 in Table 1 has propensity $k \times (\#A) \times (\#B)$. For second-order reactions, such as this one, the k term depends on the volume and should be replaced with $k'/V(t)$, where $V(t)$ is the volume and k' is independent of volume. This change leads to the propensity $a'/V(t)$, where $a' = k' \times (\#A) \times (\#B)$ is a constant independent of volume (and hence time). For simple $V(t)$, eq 9 can be solved analytically; for example, in Arkin et al.,[5] the volume is modeled as increasing linearly. Thus, $V(t) = V_0 + ct$, which leads (by a simple integration) to the distribution

$$P(t|t_0) = \frac{a'(V(t_0) + ct)^{-a'/c-1}}{V(t_0)^{-a'/c}} \tag{10}$$

Note also that in the limit as c goes to zero, this distribution reduces to an exponential with parameter a'/V_0, as expected.

It is a straightforward operation to generate random numbers according to this distribution, using the inversion-generating method:[13,16] one calculates the cumulative distribution function F (a simple integral of P) and takes a sample U from a uniform random number generator, and then the variable $F^{-1}(U)$ has the correct distribution. For the preceeding example, the variable $R = V(t_0)[U^{-c/a'} - 1]/c$ is distributed according to eq 10.

4.1.3. Generating Fewer Random Numbers. It was remarkably simple in the time-independent, exponential case to re-use the same random numbers. The extension is somewhat more difficult. The method presented here works for re-using random variables generated by the inversion-generating method. Of course, not every random variable is generated that way, and in practice it may be hard to re-use other random numbers.

Theorem 2. *Let τ be a random number generated according to an arbitrary distribution with parameter a_n and distribution $F_{a,n}$. Suppose the current simulation time is t_n, and the new parameter (after the update in Step 2) is a_{n+1}. Then, the transformation $\tau' = F^{-1}_{a,n+1}([F_{a,n}(\tau) - F_{a,n}(t_n)]/[1 - F_{a,n}(t_n)])$ generates a new random variable from the correct (new) distribution.*

The proof of this theorem is in the *Appendix*. Here are some examples:

Example 2. *For exponentials,*

$$F_{a,n}(u) = \Pr(T_n \le u) = \begin{cases} 1 - \exp(-a_n(u - t_{n-1})) & \text{if } u > t_{n-1} \\ 0 & \text{otherwise} \end{cases}$$

and

$$F^{-1}_{a,n}(U) = \begin{cases} -\ln(1 - U)/a_n + t_{n-1} & \text{if } 0 \le U \le 1 \\ \text{undefined} & \text{otherwise} \end{cases}$$

So, by the theorem,

$$\begin{aligned} \tau' &= F^{-1}_{a,n+1}([F_{a,n}(\tau) - F_{a,n}(t_n)]/[1 - F_{a,n}(t_n)]) \\ &= \frac{-1}{a_{n+1}} \ln\left(1 - \frac{[1 - \exp(-a_n(\tau - t_{n-1}))] - [1 - \exp(-a_n(t_n - t_{n-1}))]}{1 - [1 - \exp(-a_n(t_n - t_{n-1}))]}\right) + t_n \\ &= (a_n/a_{n+1})(\tau - t_n) + t_n \end{aligned}$$

This is the transformation used by the Next Reaction Method.

Example 3. *The previous section considered a process with*

$$V(t) = V(t_{n-1}) + c(t - t_{n-1})$$

$$F_{a,n}(u) = 1 - \left(\frac{V(t)}{V(t_{n-1})}\right)^{-a_n/c}$$

and

$$F^{-1}_{a,n}(U) = V(t_{n-1})[(1 - U)^{-c/a_n} - 1]/c + t_{n-1}$$

By the theorem, one can re-use random numbers by the transformation

$$\begin{aligned} \tau' &= F^{-1}_{a,n+1}([F_{a,n}(\tau) - F_{a,n}(t_n)]/[1 - F_{a,n}(t_n)]) \\ &= \frac{V(t_n)}{c}\left[\left(1 - \frac{\left(1 - \left(\frac{V(\tau)}{V(t_{n-1})}\right)^{-a_n/c}\right) - \left(1 - \left(\frac{V(t_n)}{V(t_{n-1})}\right)^{-a_n/c}\right)}{\left(\frac{V(t_n)}{V(t_{n-1})}\right)^{-a_n/c}}\right)^{-c/a_{n+1}} - 1\right] + t_n \\ &= \frac{V(t_n)}{c}\left[\left(\frac{V(\tau)}{V(t_n)}\right)^{a_n/a_{n+1}} - 1\right] + t_n \end{aligned}$$

In some cases, as in the examples just presented, it is possible to calculate a closed-form solution of the equations that is relatively simple; and, this method is practical. In general, it may not be at all practical and it may be easier to generate fresh random numbers.

4.2. Non-Markov Processes. Even though elementary reactions are Markov (i.e., do not depend on history) in the stochastic framework, it is sometimes useful to deal with non-Markov processes. For example, one may model a system using a certain set of variables for which the system is Markov or one may

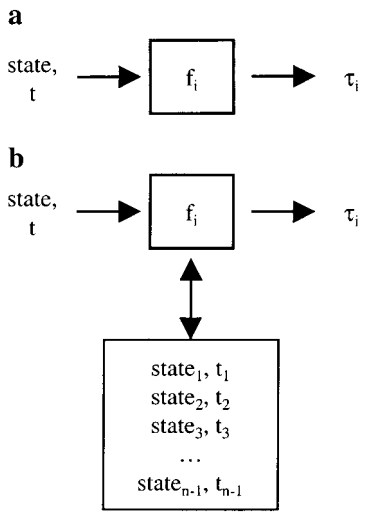


Figure 4. Procedures, *f*, for generating random numbers. (a) A Markov process, in which no history is stored. (b) A non-Markov process, which requires storing history.

use a smaller set for which the system is not. Provided the simulation algorithm still works for non-Markov processes, a smaller number of variables may be significantly faster to simulate.

For Markov processes, one generates τ_i directly from the state and the time, as in Figure 4a. For example, in the time-independent case, the value a_i is calculated from the state, and τ_i is the sum of t and a random variable with exponential distribution and parameter a_i. Notice that (1) f_i is a *random* function (i.e., calling it multiple times with the same parameters will give multiple answers), and (2) f_i is a *function* in the mathematical usage or in the computer science sense of functional programming (i.e., it does not contain any internal state). For non-Markov processes, as in Figure 4b, the distribution of τ_i depends on the history of (possibly all) states of the system from the initial time to the present. Hence, one must use a *procedure* (in the computer science sense of procedural programming) that can store previous values of the system state and time.

In general, non-Markov processes are very difficult to handle.[13] The distribution of next states or of transition times to the next state may depend on the entire history of the system. Fortunately, the sort of non-Markov processes that occur in chemical reaction simulations have some nice properties that make dealing with them easier. First, the complete history of the system is uniquely determined by the series of *discrete* transitions and transition times. Given the transitions and transition times, the state at any time t is the same as the state after the last transition before t. This simplification is enormous: because continuous transitions are not possible, one can hope to store the entire state history. Second, one may not even need the entire history. For any given reaction μ, one only needs to store that fraction of the history that affects (in the dependency graph sense) the reaction μ, which for systems with many reactions, leads to another significant reduction in the amount of storage.

For *arbitrary* non-Markov chemical reaction models, even this reduction in storage may not be enough. It may be very difficult to generate τ_i given the appropriate subset of history. In that case, it may be preferable to include the full gamut of variables to make the system Markov. In those special cases where it *is* possible to generate τ_i, one may achieve a substantial performance improvement. An example follows.

4.2.1. Example: Gamma Distribution. Consider the set of equations

$$A + B \xrightarrow{k_1} S_0$$

$$S_0 \xrightarrow{k} S_1$$

$$S_1 \xrightarrow{k} S_2$$

$$\cdots$$

$$S_{n-1} \xrightarrow{k} S_n$$

$$S_n \xrightarrow{k_2} C + D$$

Systems of equations very much like this come up in the Arkin et al.[5] model of lambda phage, both for transcription and for translation. Physically, this means that at some time a molecule of type S_0 is produced and then undergoes an n-step process. Subsequently, the resulting molecule, S_n, affects the rest of the system.

The first and last equations are different, but all the n intervening equations are identical. Assuming time independence (as is the case in the model; first-order reactions are not affected by change in volume), one may solve these n equations analytically. Rather than n exponentials, the combined waiting time is a gamma distribution. Specifically, consider a single molecule of S_0, produced at t_0, with no other molecules of S_0 produced. Then,

$$\Pr(\text{one molecule of } S_n \text{ is produced between } t \text{ and } t + dt \mid \text{one molecule of } S_0, t_0) = \frac{k[k(t - t_0)]^{n-1}}{(n-1)!} \exp[-k(t - t_0)] \times dt \quad (11)$$

This equation is simply a gamma distribution, and there are efficient ways to generate random numbers according to this distribution.

Now consider several molecules undergoing this process. This composite system can be described by ordered pairs of the form (molecule identity, state). There are two ways to simplify this system: grouping by state and grouping by molecule identity. Thus far, the grouping has always been by state (i.e., the number of molecules in state S_0, the number in state molecules of S_1, etc.). For the context in which this problem occurs, with many more states than molecules, one achieves a smaller system by grouping by molecule identity. Using eq 11, one can simplify the n-step exponential process into a 1-step gamma process; thus, the number of processes to consider is equal to the number of distinct molecules, which is much less than the number of states. (Note that this simplification is possible because the reactions involved are first-order.) These two possibilities are shown schematically in Figure 5.

The procedure f_i is as follows: rather than store state and time directly, f_i will keep a list L of processed values τ'. Every time a new molecule of S_0 is produced, f_i generates a τ' value for it according to eq 11 and adds that value to L. The value of τ_i' that f_i returns is simply the minimum of the τ values in L. (The astute reader will note that the operations required on L are insert, delete, and minimum, so one could implement L as a priority queue. This implementation of L as a priority queue should not be confused with the indexed priority queue in Section 3.2.)

TABLE 2: Key Reactions of the Model of the Bacteriophage Lambda

no.	reaction	condition
1	$RNAP_{free} + DNA_{free} \xrightarrow{k_1} RNAP{\bullet}DNA_{open,\ 0}$	
2	$RNAP{\bullet}DNA_{open,\ n} \xrightarrow{k_2} RNAP{\bullet}DNA_{open,\ n+1}$	for $n = 0...DMAX - 1$
3	$RNAP{\bullet}DNA_{open,\ DMAX} \xrightarrow{k_3} RNAP_{free} + DNA_{free} + RNA_{free}$	
4	$RNase + RNA_{free} \xrightarrow{k_4} RNase$	
5	$ribosome + RNA_{free} \xrightarrow{k_5} ribosome{\bullet}RNA_0$	
6	$ribosome{\bullet}RNA_n \xrightarrow{k_6} ribosome{\bullet}RNA_{n+1}$	for $n = 0...RMAX - 1$
7	$ribosome{\bullet}RNA_{RMAX} \xrightarrow{k_7} ribosome_{free} + RNA_{free} + protein$	
8	$protein \xrightarrow{k_8} no\ protein$	
9	$protein + protein \xrightarrow{k_9} protein{\bullet}protein$	
10	$protein{\bullet}protein \xrightarrow{k_{10}} protein + protein$	

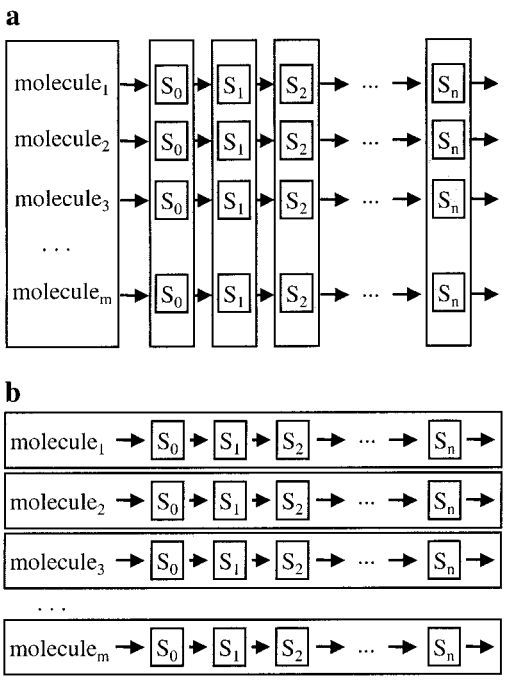


Figure 5. Ways of splitting reactions. (a) By state, (b) by molecule.

In the current context, the transcription or translation lengths, and hence the n values, may be in the hundreds or even thousands, so this enhancement achieves quite a speedup (see Section 5).

5. Application

The Arkin et al.[5] model of the bacteriophage lambda provides a good, large test case for the scalability of stochastic simulation algorithms. The details of the model can be found in the original paper; this paper shall just sketch out the model and focus on how the algorithmic improvements described apply to it.

5.1. The Model. The model consists of the complete set of equations governing the regulation of five genes in the temperate bacteriophage lambda. The key reactions are summarized in Table 2. The model considers five genes—*N, cro, cI, cII,* and *cIII*—and their protein products. There are five sets of equations similar to those in Table 2. Two of the five genes have protein products that dimerize; that is, can exist as a protein$_2$ form, as in Reactions 9 and 10 in Table 2. The other three proteins exist only as monomers, so Reactions 9 and 10 do not apply to those proteins.

TABLE 3: Performance Data

a. Next Reaction Method	
reactions	78
average out-degree	4.2
trajectories generated	500
simulation events	35 000 000
updates	206 000 000
operations	760 000 000
operations/event	22
operations/update	3.7
b. Gamma Distribution	
gamma reactions	21
equivalent elementary reactions	10 000
gamma simulation events	380 000
equivalent elementary events	99 000 000

The key phenomenon in gene regulation is that the expression of certain genes regulates the expression of other genes: the "constants" k_i in the equations may depend in some way on the number of proteins present. This extra dependency is the sort that was alluded to in the definition of *DependsOn*(a_μ). For the lambda model in particular, the k_1s can be calculated from the concentrations of the various proteins using a straightforward equilbrium thermodynamics model. See Arkin et al.[5] or Shea et al.[17] for a complete description of how to calculate the k_1 values.

In addition to the equations shown, Arkin et al.[5] provide a more detailed model of degradation for two of the proteins and also a more detailed model of transcriptional termination. Also, these reactions take place in a growing *Escherichia coli* host cell, so they model the volume increase over time as well (see the discussion in Section 4.1.2).

5.2. Results. Table 3a shows performance data for the Next Reaction Method, with all the enhancements of Section 4, on the (full) Arkin et al.[5] model. Note that the average out-degree — the average number of edges from a given vertex in the dependency graph — is significantly lower than the number of reactions, so the dependency graph is sparse, as mentioned in Section 3.1. Five hundred trajectories were generated for the condition "1 phage per cell"; each required ~70 000 simulation events.

Based on the number of simulation events and the average out-degree, one would expect 150 million updates (executions of Step 5), or one update per out-edge per simulation event. The actual number is substantially higher at 210 million because certain reactions are executed much more often than others. Under the conditions used (one phage per host cell), the two

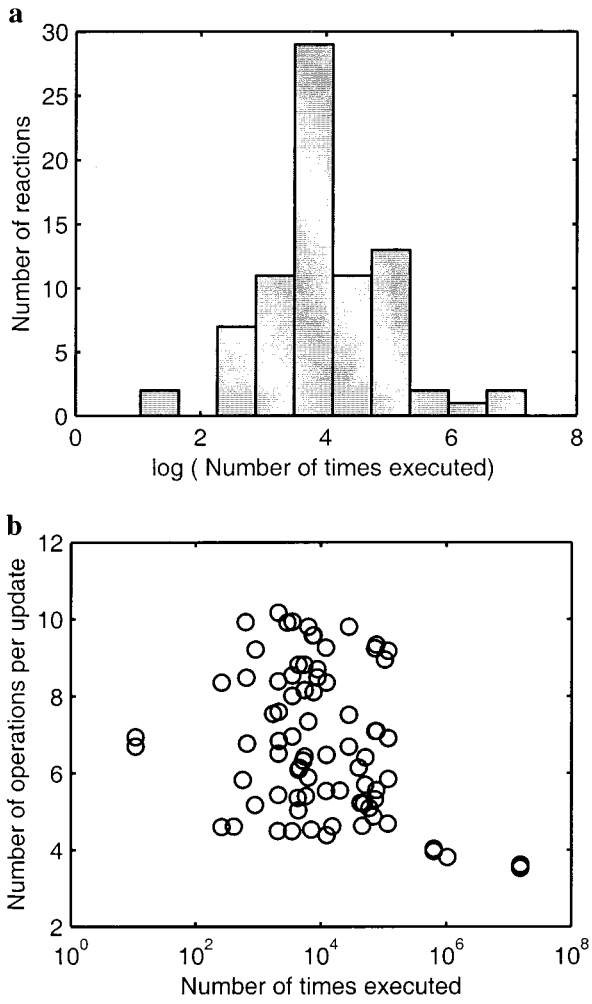


Figure 6. Performance data for lambda model. (a) Number of times each reaction is executed. (b) Number of update operations per executed reaction.

most frequent reactions constitute nearly 90% of all simulation events. Both of these most-frequent reactions have out-degree 6, which explains why the number of updates per simulation event is ~6, not 4.2 as would be expected.

Which reactions are executed most often? By definition, those reactions with high propensities are more likely to be executed. Unfortunately, because propensity is a function of rate constant *and of state*, it is very difficult to calculate which reactions will occur most frequently without just running the whole simulation. (For example, by changing the conditions to "six phages per host cell", with all the same rate constants, two additional reactions occur with approximately the same frequency as the original two. Under these conditions, these four most likely reactions are responsible for 95% of the the total of 180 million simulation events.) It is not obvious whether there is a way to identify such reactions and deal with them separately (e.g., by separation of time scales) *without affecting the results*.

Fortunately, as Figure 6b shows, the optimizations mentioned have the effect that the most frequently executed reactions are also the most efficient. Note that each time a reaction is executed and a new τ_i has to bubble down, at least two comparison operations are required, namely, comparing with the left and right children. (Bubbling up only requires a single comparison.) The fact that the most frequently executed reactions have efficiencies around 3 and the overall total is 3.7 means that the updated τ_is do not typically move very far from their original positions in the indexed priority queue, as mentioned in Section 3.2.

TABLE 4: Comparison of Methods for the Lambda Model[a]

parameter	direct method	optimized direct method	next reaction method
a_i calculations	2700	210	210
Other $\times$, $\div$	0	0	340
$+$, $-$, comparison	2900	1500	1100
exp random numbers	35	35	35
uniform random numbers	35	35	0

[a] All numbers are in millions.

Also notice that by re-using random numbers, the number of calls to the pseudo random number generator is approximately equal to the number of simulation events, not to the number of updates. This re-use decreases the number of random number generator calls sixfold from the Absolute Time First Reaction Method.[4] Equivalently, one can simulate six times as much without having to worry about numerical problems.

Another big time savings comes from using the gamma distribution, as detailed in Table 3b for the condition "one phage per host cell". One could, in principle, write the composite reactions as their elementary steps; however, there would be nearly 500 times more reactions under that scheme. Furthermore, that scheme would involve 260 times the number of simulation events — nearly three times the *total* number of simulation events — just for the gamma reactions. The total simulation would take four times as long without this particular refinement. Thus, using the gamma distribution provides enormous savings: all 21 gamma reactions together were only 1% of the simulation events.

Table 4 compares a simple implementation of Gillespie's Direct Method with the optimized version of the Direct Method in the *Appendix* and with our Next Reaction Method for the lambda model. All methods include the gamma optimization (note that said optimization is possible for both the Direct and Next Reaction methods, but slightly easier to implement in the Next Reaction Method because one already has the queuing structure in place). All numbers are rounded. (For the analysis that follows, we shall treat each "gamma reaction" just like any other reaction, rather than doing any detailed analysis of these special cases. Because those reactions constitute 1% of the total, and because $\log(78)/\log(78-21)$ is within 10% of unity, we shall not introduce huge errors.) The simplest implementation of the Direct Method updates all a_is each iteration. The Optimized Direct Method and the Next Reaction Method use dependency graphs (Section 3.1) to minimize the number of updates. The Direct Method requires 78 additions per iteration to calculate the cumulative sums a_1, $a_1 + a_2$, $a_1 + a_2 + a_3$, ..., $a_1 + a_2 + a_3 + ... + a_{78}$, and log 78 comparisons to do a binary search and find the appropriate μ. The Optimized Direct Method takes log 78 additions per *update* and log 78 comparisons per iteration to generate μ. The Next Reaction Method requires an addition, a subtraction, a multiplication, and a division per reused τ, of which there are #updates − #events, and it also requires 2 log(78) operations per simulation event *in the worst case*, but the efficient UPDATE algorithm (Algorithm 3) may cut down on that number substantially. Finally, recall that the Direct Method uses two random numbers per simulation event, whereas the Next Reaction Method uses one. Different operations take different amounts of time, so Table 4 does not attempt to combine all operations into a single number.

As can be seen from the table, the Optimized Direct Method and the Next Reaction Method both outperform the Direct Method. The relative performances of the Optimized Direct Method and the Next Reaction Method are hard to compare quantitatively without more specific data about speeds of various steps but, qualitatively, the Next Reaction Method will evidently outperform the Optimized Direct Method.

6. Conclusions

This paper presents an exact, efficient way to do calculations for large systems of loosely coupled reactions in the stochastic framework. The algorithms presented are *exact* (i.e., provably equivalent to the chemical Master Equation approach) and they are *efficient*, both in running time and in number of random numbers generated. For both the Direct and the Next Reaction Method, the amount of time required per iteration is proportional to the logarithm of the number of reactions, not the number itself. The number of random numbers generated by the Next Reaction Method is *(reactions) + (simulation events)*, which is only slightly higher than the optimal number, *simulation events.* (For comparison, the Direct Method takes $2 \times$ *simulation events*.)

The paper extends the Next Reaction Method to time-varying rate constants while maintaining both types of efficiency. As an aside, the *Appendix* shows how to extend the Direct Method to time-varying rate constants, but it is not clear whether this is useful.

The paper also provides one specific example of extending of the Next Reaction Method to non-Markov processes and sketches a framework in which other examples might be formulated.

For loosely coupled chemical reaction systems in solution, the Next Reaction Method is preferable to the Direct Method for the following reasons:

- although the efficiency of updates in each is $O(\log r)$, if some reactions are much faster than others, the Next Reaction Method may be effectively $O(\log r')$, where $r' \ll r$;
- the Next Reaction Method can easily be enhanced to re-use random-numbers, which reduces the number of random numbers to half as many as used by the Direct Method; and
- the Next Reaction Method is easily extended to both Markov and non-Markov time-varying processes.

Finally, the paper presents an example of the performance of these algorithmic improvements on a test case from the biology literature.

Acknowledgment. The authors thank Daniel Gillespie and Tau-Mu Yi for their careful reading of an earlier version of this paper. We also appreciate several comments from the anonymous reviewers that improved the clarity of this paper. Supported in part by ONR grant N00014-97-1-0293, by a JPL-CISM grant, by NSF Young Investigator Award CCR-9457811, and by a Sloan Research Fellowship.

7. Appendix

7.1. Proof of Theorem 2. The random number τ is orginally distributed according to distribution $F_{a,n}$, with density $P_{a,n}$. After Step 6, it is distributed according $P'_{a,n} = P_{a,n}(u|T > t_n)$, which is equal to

$$P'_{a,n}(u) = \begin{cases} P_{a,n}(u)/[1 - F_{a,n}(t_n)] & \text{if } u > t_n \\ 0 & \text{otherwise} \end{cases}$$

By the Random Variable Transform Theorem,[13] the random variable $Y = [F_{a,n}(\tau) - F_{a,n}(t_n)]/[1 - F_{a,n}(t_n)]$ has the density

$$\begin{aligned} Q(y) &= \int_{-\infty}^{\infty} P'_{a,n}(u)\delta\left(y - \frac{F_{a,n}(u) - F_{a,n}(t_n)}{1 - F_{a,n}(t_n)}\right) \mathrm{d}u \\ &= \frac{1}{1 - F_{a,n}(t_n)} \int_{t_n}^{\infty} P_{a,n}(u)\delta\left(y - \frac{F_{a,n}(u) - F_{a,n}(t_n)}{1 - F_{a,n}(t_n)}\right) \mathrm{d}u \\ &= \frac{1}{1 - F_{a,n}(t_n)} \int_{F_{a,n}(t_n)}^{F_{a,n}(\infty)=1} \delta\left(y - \frac{v - F_{a,n}(t_n)}{1 - F_{a,n}(t_n)}\right) \mathrm{d}v \\ &= \int_0^1 \delta(y - w)\, \mathrm{d}w \\ &= \begin{cases} 1 & 0 < y \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The second line is just the definition of $P'_{a,n}(u)$. The third line comes from the transformation $v = F_{a,n}(u)$, $\mathrm{d}v = [\mathrm{d}F_{a,n}(u)/\mathrm{d}u]\mathrm{d}u = P_{a,n}(u)\mathrm{d}u$. The fourth line comes from the transformation $w = [v - F_{a,n}(t_n)]/[1 - F_{a,n}(t_n)]$, $\mathrm{d}w = (1/[1 - F_{a,n}(t_n)])\mathrm{d}v$. The final line comes from the definition of the delta function.

Hence, the random variable Y is distributed uniformly on (0, 1]. Finally, the inverse generation method works by transforming a uniform random number U to $F^{-1}(U)$. Here, Y is such a uniform random number, which proves the theorem.

7.2. Enhancing the Direct Method. The underlying ideas of the Next Reaction Method — use a dependency graph to update the minimal number of variables and use an efficient data structure — can be applied to the Direct Method as well.

Algorithm 6. *(Efficient Exact Stochastic Simulation — Direct Method), Replace Steps 1 and 2 of the Direct Method with:*

1. *Initialize (i.e., set initial numbers of molecules, set $t \leftarrow 0$, generate a dependency graph G).*
2. *Calculate the propensity function, a_i, for the following i:*

   If this is the initial iteration, calculate a_i for all i;

   Otherwise, let μ be the reaction that was just executed. For each edge (μ,α) in the dependency graph G, update a_α.

It is clear from the definition of the dependency graph that this will update only the a_is that need to be updated. The only thing remaining is to use the right data structure to speed up the updates.

To complete the speed up of the Direct Method, one must do Steps 2, 3, and 4 efficiently. One might consider using a simple array to store each of the a_is. In this scheme, updates would be very fast. However, Step 3 of the Direct Method would then take time proportional to the number of reactions. A better data structure, which takes time proportional to the logarithm of the number of reactions, follows.

Store the a_is as the leaves of a complete tree and store in each non-leaf node the sum of its left child and right child (see Figure 7a). Thus, the root will have value $\sum_i a_i$. When a_is change, update (1) those a_is that have changed and (2) their ancestors. Notice that the tree contains r leaves and $r/2 + r/4 + r/8 + \ldots + 1 \cong r$ non-leaves. The height of the tree is simply $\log 2r = 1 + \log r$. Each update affects one node at each level, hence is $O(\log r)$.

Generating the random numbers τ will be easy because the root of the tree contains the appropriate parameter. For the μ

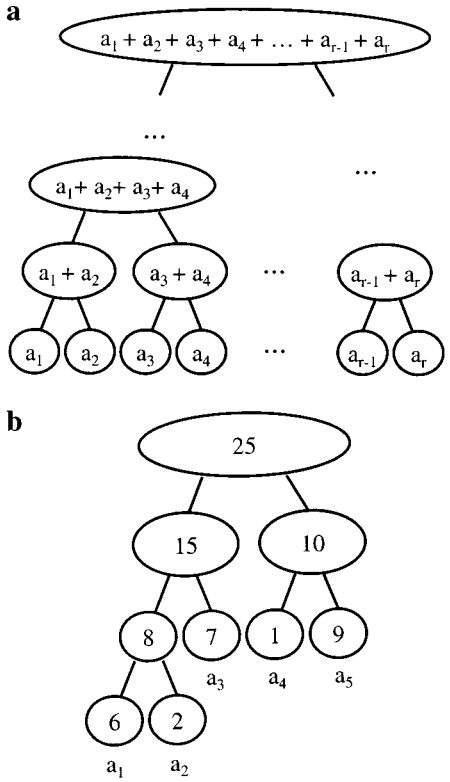


Figure 7. Data structure used for a_is for an efficient version of the Direct Method. Each leaf contains an a_i value. Each other node contains the sum of its left and right child. (a) Generic construction, (b) numerical example used in text.

value, generate a random number x between 0 and $\sum_i a_i$ (which can be found at the root) and then use the following algorithm, starting at the root.

Algorithm 7. (Efficient Uniform Random Number Generation).

1. If the current node is a leaf, let μ be its index.

2. Otherwise, if $0 \leq x \leq$ (left child value), then call this algorithm recursively on the left child with parameter x.

3. Otherwise, (left child value) $\leq x$; so call this algorithm recursively on the right child with parameter $x -$ (left child value).

The discussion thus far has been somewhat abstract and calls for an example.

Example 4. *Consider the numerical values in Figure 7b, the tree structure for $a_1 = 6$, $a_2 = 2$, $a_3 = 7$, $a_4 = 1$ and $a_5 = 9$. To calculate τ, generate an exponential random variable with parameter 25. For the μ value, generate a random number x between 0 and 25. Suppose $x = 15.5$. Because $x > 15$ (the left child), go to Step 3 of the algorithm and descend the right subtree (i.e., the one whose root is "10") with parameter $x' = x - 15 = 0.5$. Now, $0 \leq x' \leq 1$, so use Step 2 of the algorithm and go to the left node, labeled "1". Finally, this is the leaf corresponding to a_4, so stop and let μ be "4".*

One standard method of generating a random number of this distribution is to generate a random number R between 0 and 1, then find the index μ such that $\sum_{i=1}^{\mu-1} a_i \leq R \times (\sum_i a_i) < \sum_{i=1}^{\mu} a_i$. In fact, Algorithm 7 does precisely that, in an efficient way, and takes time proportional to the height of the tree, not the total number of nodes in the tree.

With this algorithm, an update takes $1 + \log r$ operations, and generating μ takes $\log r$ operations. By the sparseness assumption, each simulation event (time through the loop) takes at most k $(1 + \log r)$ operations, where k is a constant independent of r. For E simulation events, the algorithm takes $O(E \log r)$ operations, not counting the initialization in Step 1.

As a side note, there are other efficient ways to generate random variates of a discrete distribution,[18] which are somewhat esoteric but have better expected times. One could do a thorough analysis of the trade-off between programming complexity, run time, numerical stability, and number of uniform random numbers required. However, because the Next Reaction Method is more easily enhanced to use fewer uniform random numbers and handle time-varying processes, both Markov and non-Markov, we shall favor it.

7.3. Time-Varying Direct Method, Markov Processes. It is well known how to generalize the Direct Method to arbitrary functions of time $a_i(t)$.[13,19] One writes an equation that is analogous to eq 4. If S is the state at time t_0, then the equation is

$$P(\mu,\tau|S,t_0) = a_\mu(S,\tau) \exp\left(-\int_{t_0}^{\tau} \sum_j a_j(S,t)\mathrm{d}t\right)$$

At each subsequent step of the algorithm, one must recondition; that is, change the density $P(\mu,\tau|S,t_i)$ to the density $P(\mu,\tau|S,t_{i+1})$. Reconditioning works out to changing the lower limit of integration.

It may be hard to generate random numbers according to this distribution for arbitrary functions of time a_i. (Note, in particular, that the lower limit of integration changes each iteration, so methods that involve numerical storage of partial values of the integral will have to do significant recalculation each iteration.) If all the a_is change *in the same way*, then one can use the enhancements of the previous section; if not, it is not immediately clear how to run this algorithm efficiently for many reaction channels. Once again, the Next Reaction Method is preferable.

References and Notes

(1) Gibson, M. A.; Mjolsness, E.; *Computational Methods for Modeling Biochemical Networks*; Bower, J., Bolouri, H., Eds.; MIT Press: Cambridge, MA, in press.

(2) Gillespie, D. T. *Physica A* **1992**, *188*, 404.

(3) Gillespie, D. T. *J. Phys. Chem.* **1977**, *81*, 2340.

(4) Lukkien, J. J.; Segers, J. P. L.; Hilbers, P. A. J.; Gelten, R. J.; Jansen, A. P. J. *Phys. Rev. E* **1998**, *58*, 2598.

(5) Arkin, A. P.; Ross, J.; McAdams, H. H. *Genetics* **1998**, *149*, 1633.

(6) van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*; Elsevier: Amsterdam, The Netherlands, 1992.

(7) Morton-Firth, C. J. *Stochastic Simulation of Cell Signalling Pathways*, Thesis, University of Cambridge, 1998.

(8) Colquhoun, D.; Hawkes, A. G. *Phil. Trans. R. Soc. Lond. B* **1982**, *300*, 1.

(9) Gillespie, D. T. *J. Comput. Phys.* **1976**, *22*, 403.

(10) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; The MIT Press and McGraw-Hill Book Company: Cambridge, MA and New York, NY, 1990; Chapter 7.

(11) It may happen that $a_\alpha = 0$ for some $\alpha \neq \mu$, in which case Step 5b is incomplete. As long as $a_\alpha = 0$, τ_0 should be set to ∞. Let t_1 be the time at which a_α first becomes 0, let t_2 be the time at which a_α ceases to be 0, let $a_{\alpha,old}$ be the last pre-0 propensity, and let $a_{\alpha,new}$ be the first post-0 propensity. Then, the correct transformation for Step 5b is $\tau_\alpha \leftarrow (a_{\alpha,old}/a_{a,new})(\tau_\alpha - t_1) + t_2$.

(12) Segers, J. *Algorithms for the Simulation of Surface Processes*, Thesis, Eindhoven University of Technology, 1999.

(13) Gillespie, D. T. *Markov Processes: An Introduction for Physical Scientists*; Academic: San Diego, CA, 1992.

(14) Similarly, plugging into the RVT theorem shows directly that the formula for the case $a_i = 0$ produces the the correct new distribution for T_i'. We shall outline a more intuitive argument, but not make it entirely rigorous: Intuitively, one can multiply the original transformation by $a_{i,n+1}$, giving $a_{i,n+1}\tau_i' = a_{i,n}\tau_i - a_{i,n}t_{n+1} + a_{i,n+1}t_{n+1}$. One can view this as a transformation $r' = r - a_{i,n}t_{n+1} + a_{i,n+1}t_{n+1}$ of a new random variable r,

which is well defined even if $a = 0$. Applying the $r \rightarrow r'$ transformation for each time that $a_{i,n+1} = 0$ and the first non-zero $a_{i,n+l}$, and applying $r = a_{i,n}\tau_i$ to the last pre-zero a_i and the first post-zero a_i, gives the desired transformation.

(15) For time-varying processes, there may be a non-zero probability that a given reaction does not occur at all (in other words, integrating eq 9 from t_n, to ∞ may result in some $p < 1$). In this case, one must choose random numbers in such a way that $P_\mu(\infty|S,t_n) = 1 - p$. This choice is easy to make in the Next Reaction Method, one generates a uniform random number r: if $r > p$, the reaction never occurs so $\tau = \infty$; otherwise one transforms r (or uses some other method) to find $\tau < \infty$.

(16) Leon-Garcia, A. *Probability and Random Processes for Electrical Engineering*; Addison-Wesley Publishing: Reading, MA, 1994.

(17) Shea, M. A.; Ackers, G. K. *J. Mol. Biol*. **1985**, *181*, 211.

(18) Matias, Y.; Vitter, J. S.; Ni, W. C. "Dynamic Generation of Discrete Random Variates", Technical Report, Bell Labs, 1997.

(19) Jansen, A. P. J. *Comput. Phys. Commun*. **1995**, *86*, 1.