The Microsoft Research - University of Trento
Centre for Computational and Systems Biology

Technical Report CoSBi 12/2008

Exactness and Approximation of the Stochastic Simulation Algorithm

Ivan Mura

The Microsoft Research - University of Trento
Centre for Computational and Systems Biology

mura@cosbi.eu

Exactness and Approximation of the Stochastic Simulation Algorithm

Ivan Mura

The Microsoft Research - University of Trento
Centre for Computational and Systems Biology

mura@cosbi.eu

Abstract

This short note intends to clarify about the applicability of the Stochastic Simulation Algorithm proposed by Gillespie for the analysis of systems of coupled biochemical reactions. The derivation of Gillespie's results is revisited to pinpoint those steps at which, depending on the validity of the assumptions adopted about the system to be studied, approximations may be introduced. We discuss about the ways the inaccuracies entailed by the approximations may propagate and affect simulation results.

1 Introduction

In 1976, a paper by Daniel T. Gillespie [5] proposed a novel computational approach to effectively analyze the time behavior of chemical/biochemical systems. That paper provided an easy to implement algorithm for simulating the evolution of a system together with a theoretical justification, grounded on statistical mechanics, of its applicability. The success of Gillespie's method for the study of biochemical systems dynamics is well demonstrated by the plethora of studies, papers and computational tools based on it that have appeared since the original publication.

The main reasons for this widespread acceptance stem from the simplicity of the proposed algorithmic approach, which easily lends itself to straightforward (though non necessarily optimized) implementations, and from the clear link that is maintained with the intuitive descriptive language of chemical reactions (which also suits biochemistry). In fact, Gillespie's algorithm can be seen as a formalization of the common intuitive understanding of how a chemical or a biochemical system described through chemical reactions would evolve over time.

On the other hand, Gillespie also postulated that a very specific choice about the probability distribution of the times at which reactions occur leads indeed to an accurate picture of the system dynamics, and provided theoretical justification for such a choice. This justification is valid under a precise set of assumptions. Notably, for this set of assumptions to be valid, the system does not necessarily need to approach the thermodynamic limit. Therefore, the main Gillespie's result still applies to systems composed of few chemically interacting molecules.

Nonetheless, the assumptions on which Gillespie based his results may not be trivially valid for many systems of interest in biology. In fact, these assumptions should be whenever possible confirmed or rejected, and in any case questioned and not taken a priori. In this respect, there is a myriad of examples of studies assuming the general validity of such assumptions without discussion.

This note does not aim at extending Gillespie's results nor their applicability, which have been already consistently described by Gillespie himself in his papers [5, 6]. Rather, the objective is to focus on the way the assumptions made about the system imply the validity of the Gillespie's approach to stochastic modeling and simulation, as well as to reason on which approximations are introduced when such assumptions are not valid and to discuss about the ways the inaccuracies entailed by the approximations may propagate and affect the results. To this, the first two published papers of

Gillespie [5, 6] are revisited, with the objective of extrapolating and putting in a system modeling perspective the main results contained therein. The adopted modeling perspective helps in elucidating the role of assumptions and their consequences.

The rest of this document is organized as follows. In Section 2 we outline the main contributions proposed by Gillespie and describe the importance of the SSA as an effective computational tool for the simulation of biochemical systems. Section 3 presents the set of hypotheses that Gillespie considered for ensuring the validity of its stochastic modeling approach. In Section 4 we discuss on cases when such hypotheses are not trivially satisfied, and we point out where approximations may be introduced. The potential impact of those approximations is shortly considered in Section 5. Finally, conclusions are provided in Section 6.

2 Gillespie's results

The material presented in Gillespie's paper of 1976 [5] and 1977 [6] provides a very clear description of the theoretical basis upon which the stochastic description of system dynamics is built and the stochastic simulation algorithm (SSA, hereafter) is formulated. Further papers from Gillespie and many other authors focused on the optimization of the SSA [4, 2], on its approximate versions [8] and on the relationships between the stochastic and deterministic modeling approaches [7].

In fact, the results that can be identified in Gillespie's first two papers rely on a precise characterization of the stochastic behavior of chemical systems in terms of the probability of the occurrence of a reaction. In the following, the state of a chemical system is a vector whose components represent the number of molecules of each chemical species, denoted by $\vec{x}$s. Obviously, $\{x\}_t$, $t \geq 0$ is a stochastic jump-process, which probabilistically moves from one state to the other. At the basis of Gillespie's approach is the following assumption (which Gillespie calls the *fundamental hypothesis*):

Hypothesis 1. *For every reaction j in the system, if $\vec{x}$ is the state of the system at time t, $t \geq 0$, the probability that the a reaction of type j occurs in the next infinitesimal time interval $t + \Delta t$ can be expressed as $a_j(\vec{x}) \cdot \Delta t$, where $a_j(\vec{x})$ does not have any explicit dependency on t.*

Functions $a_j(\vec{x})$ are called *propensity* functions. Then, Gillespie's papers present the two following major contributions:

- the proof that, under a precise set of assumptions, the fundamental hypothesis holds for chemical systems;

- the definition of the Stochastic Simulation Algorithm (SSA, hereafter), to produce sample realizations of the stochastic processes $\{x\}_t$ underlying chemical systems for which the fundamental hypothesis is valid.

The first result proves the validity of the fundamental hypothesis of Gillespie by introducing a set of sufficient conditions at the physical molecular level. These conditions, which will be reviewed in detail in the next section, are easily verified when the interacting species form a gas, where reactant molecules undergo many nonreactive collisions. Occasionally, a collision involves two molecules that are able to react, and a reaction may take place. Because of the particular state of a gas, the reaction products diffuse very quickly and the system maintains homogeneity.

The fundamental hypothesis may also hold of reactions happening in other types of systems, as it seems to be proved by the succesful validation of many stochastic models of biochemical systems against experimental data. However, Gillespie's results only allow us taking it for granted for systems that are well-mixed, under thermal equilibrium, and only for specific types of chemical reactions, often termed *elementary*. In this context, an elementary reaction is one that does not abstract any intermediate species. There are only two types of elementary reactions, namely:

1. the *monomolecular* reaction, where a molecule of species A transforms itself into a set of reaction product molecules;

2. the *bimolecular* reaction, where a molecule of species A and a molecule of species B, where species B may be the same as species A, bind to generate a new molecule.

Examples of reactions of type 1) reactions include the isomerization $A \to B$, and the split of a molecule into sub-molecules $A \to B + C$. Examples of type 2) reactions include complexation $A+B \to C$, and, as a special case, the dimerization $A+A \to A_2$. Any reaction that involves more than two reactant molecules is not elementary. A simple statistical mechanics argument shows that the likelihood of more than two molecules simultaneously colliding in a reaction vessel is infinitesimal, and any reaction involving more than two reactant molecules, such as $A+B+C \to D$, must occur indeed as a sequence of bimolecular reactions, for instance $A + B \to E$, $E + C \to D$.

The reason why it is so desirable that the fundamental hypothesis holds of a chemical system is that it implies that, in every state of the system

and for every reaction j the time to the next occurrence of reaction j is a random variable following a negative exponential distribution. This allows describing the evolution of the system over time from an initial state, through a simple system of first order linear differential equations, known as the Chemical Master Equation (CME, hereafter). In probability theory, the CME describes the evolution of a Continuous-Time Markov Chain (CTMC, hereafter), whose state space corresponds to the set of possible states of the chemical system, and whose transitions correspond to the occurrence of reactions. A rich set of exact analysis techniques and numerical solution approaches for both transient (time-dependent) and steady-state analysis of CTMC exists [3], which can be exploited to investigate the dynamic evolution of a biochemical system satisfying the fundamental hypothesis 1).

In fact, the second contribution offered by Gillespie, i.e. the SSA algorithm, is a simulation scheme for generating realizations of the CTMC. The SSA scheme has given rise to a family of computationally efficient algorithms for simulating CTMCs that represent the evolution of a set of coupled biochemical reactions. These algorithms [4, 6, 2] are much more efficient than general event-driven simulation algorithms, which also apply to CTMCs.

The SSA is exact, in the sense that it generates only possible realizations of a CTMCs with correct probability. The exactness we are talking about here has in fact nothing to do with whether the fundamental hypothesis 1) is valid or not for a biochemical system. Assuming that the hypothesis 1) holds of a biochemical system described as a set of reactions, the SSA provides a way to explore the dynamic evolution of the system, without requiring any additional assumption and without introducing any further approximation.

For the sake of completeness, we have to mention that, as in any stochastic simulation approach, the results SSA can provide are in the end approximate because the number of distinct realizations of a CTMC is infinite, and the exact evaluation of the measures of interest would require including the contribution of each of them. This is obviously infeasible, and thus only approximations of the measures can be estimated. However, it is important to remark that this is not a limit specific to the SSA algorithm, but rather a common drawback of stochastic simulation, and that furthermore the quality of the approximation can be improved at the expense of a higher computation cost.

In the following section we will detail on the conditions that Gillespie introduced as sufficient to guarantee the validity of the fundamental hypothesis 1), and on where each of them contributes to back it up.

3 Sufficient conditions for the validity of the fundamental hypothesis

Gillespie took care of demonstrating that the fundamental hypothesis is indeed satisfied by some non-trivial chemical systems. He did not define directly the borders of their applicability, but rather identified the a set of sufficient conditions that ensure the validity of his results.

Specifically, Gillespie introduced the following two hypotheses:

Hypothesis 2. *The chemical system is under thermal equilibrium conditions.*

Hypothesis 3. *The chemical system is such that, at any time t, the concentration of each species is homogeneous in the reaction vessel (i.e. does not depend on space).*

It is important to notice that the homogeneity described in hypothesis 3) is in fact achieved if non-reactive collisions are much more frequent than reactive ones, which ensures diffusion processes proceed at much higher rate than any reaction in the system. Another major hypothesis was introduced for bimolecular reactions:

Hypothesis 4. *In a bimolecular reaction, the time to the occurrence of the reaction is largely determined by the time to the reactive collision, whereas the time necessary for the chemical transformation of the colliding species into the reaction products is negligible.*

If a system satisfies hypotheses 2), 3) and 4) and its reactions are only elementary ones, then we can rely on Gillespie's results, which proves that the fundamental hypothesis 1) is satisfied, the same as to say that the evolution of the state of the system over time is described by a CTMC whose transition rates from any state $\vec{x}$ are given by the propensity functions $a_j(\vec{x})$, for any j.

Gillespie actually computed the propensity functions $a_j(\vec{x})$ for elementary reactions, assuming that the molecules are approximately spherical. For instance, when this is true, functions $a_j(\vec{x})$ for a bimolecular reaction $A + B \to C$ turns out to be as follows:

$$a_j(\vec{x}) = x_A x_B V^{-1} \pi r_{AB} \sqrt{\frac{8kT}{\pi m_{12}}}$$

where x_A and x_B are the number of molecules of A and B in current state of the system $\vec{x}$, V is the volume of the reaction vessel, r_{AB} is the distance

between the geometrical centers of two reactant molecules A and B at which the reaction happens, k is the Boltzmann's constant, T is the absolute temperature and m_{AB} the reduced mass defined as $m_{AB} = m_A m_B/(m_A + m_B)$, m_A and m_B being the mass of molecule A and B, respectively.

A similar form of the propensity function can be obtained for the dimerization reaction, whereas for a monomolecular reaction $A \rightarrow B$ the propensity reaction will be in the form $a_j(\vec{x}) = x_A c_j$ where c_j is a constant. However, if the spherical assumption is not justified, we can still assume the fundamental hypothesis is applicable, solely the exact expression of the propensity functions will have a different form.

From the material presented above it is possible to define a class of systems for which it is possible to assume, without the necessity of any further discussion, the validity of the fundamental hypothesis 1). For this class of systems, we know that the SSA algorithm of Gillespie provides a tool for the exact evaluation of system dynamics, still in the aforementioned limits of stochastic simulation. However, this does not mean at all that the fundamental hypothesis will be valid only for that class of systems. In fact, the conditions described in Gillespie paper are sufficient and not necessary, and there may be many other cases for which it holds. Still, it is important to remark that whenever the sufficient conditions are not obviously satisfied, the validity of the fundamental assumption should be questioned. In the next section we will consider some examples, commonly seen in the computational biology literature, for which such validity is, according to what we stated above, to be discussed.

4 Cases when Gillespie's hypothesis is to be verified

Biochemical systems can be seen as particular instances of chemical systems where reacting molecules are heavy biological compounds such as proteins and nucleic acids, obviously far from gaseous conditions. In the small volume of a living cell the are normally no temperature gradients, which entitles us considering hypothesis 2) applicable. Similarly, the diffusion processes in a cell are also quite efficient. Even though each type of cell uses indeed various mechanisms to regulate the concentration of molecules in different areas, it is widely accepted to consider that in small volumes homogeneity is assured. This observation would lead us to consider hypothesis 3) valid of biochemical systems in limited volume areas, for instance within a compartment.

More discussion is required for hypothesis 4). In the simple reaction

$A+B \rightarrow C$, the time of the reaction can be considered negligible with respect to the collision time only if it does not involves complex transformations, such allosteric changes, of the two binding molecules. Whereas this may be considered the case for simple molecules complexation, it may not be valid in biochemistry, where reactants may be heavy structured molecules whose binding may be just the first step of a conformational rearrangement.

In fact, the situation mentioned above is a particular example of the difficulty that can be encountered in describing a biochemical system in terms of elementary reactions. Quite often, there is an incomplete knowledge of the full set of reactions, and mesoscopic or macroscopic transformations are the only ones observable. For instance, when dealing with the set of elementary reactions

$$\begin{aligned} A + E &\rightarrow AE && (1)\\ AE &\rightarrow A + E && (2)\\ AE &\rightarrow B + E && (3) \end{aligned}$$

the Michaelis-Menten abstraction in the form $A + E \rightarrow B + E$ is commonly used. This is because the only experimentally measurable process in this accumulation of the reactions product B, whereas the speed of binding/unbinding of compound AE from/to reactant A and enzyme E is not observable.

Now, Gillespie's results tell us that the fundamental hypothesis is satisfied for the biochemical system consisting of the full set of elementary reactions in 1), refunbind) and 3). Therefore, we can easily conclude that assuming the fundamental hypothesis holds of reactions $A + E \rightarrow B + E$ is entailing an approximation. Indeed, if time to next occurrence of each elementary reactions follows a negative exponential distribution (remember this is the mathematical meaning of the fundamental hypothesis), the time to the occurrence of the abstract reaction that represents the macroscopic transformation will have a distribution that results from the composition of those of the elementary reactions. Apart for the operations of multiplication for a scalar and minimum, the class of negative exponential random variables is not closed with respect to composition. For instance, the sum and maximum of a set of negative exponential random variables is not a negative exponential random variable. Thus, we can falsify the validity of the fundamental hypothesis with a simple mathematical argument.

However, assuming the fundamental hypothesis holds of an abstract reaction can be quite a good approximation. This is for instance the case for reactions 1), refunbind) and 3). Specifically, it has been proved [8] that,

if the enzyme E is quickly saturated by the substrate A, or alternatively if the speeds of the binding reaction 1) and unbinding reaction 2) is much higher than the one of the catalysis reaction 3), the fundamental hypothesis can be considered valid and the propensity $a(\vec{x})$ of the abstract reaction $A + E \rightarrow B + E$ can be expressed through the following Michaelis-Menten equation:

$$a(\vec{x}) = \frac{V_{max} \cdot x_A}{K_m + x_A} \tag{4}$$

where x_A is the number of molecules of the substrate species A in the state $\vec{x}$, and V_{max} and K_m are two constants.

This approximation is an accurate description of the process through which species B is produced only if the conditions stated above on the rates and abundance of species are satisfied. If not, it is just an approximation that can turn out in a dynamics differing from the one of the complete system described by reactions 1), refunbind) and 3).

Still, it is important to observe that a propensity function described by equation 4) still allows using the SSA algorithm to simulate the evolution of the abstract system, and that the results of simulation will be exact, with respect to the CTMC defined by the abstract reaction. This exactness derives from the fact that the SSA does not introduce any new assumption on the system. Rather, assumptions are introduced when the validity of the fundamental hypothesis is postulated.

Consider now the process of gene transcription, which is commonly modeled by a reaction of the type $G + P \rightarrow G + T + P$, where G represents a gene, P an RNA polymerase molecule, and T an mRNA transcript molecule. Let us ask whether the fundamental hypothesis would hold of such a reaction. If one considers the complexity of the transcription process, which encompasses the sequential assembly of a long nucleotide sequence based on the gene template scan, it is quite intuitive to understand that it is not an elementary reaction. Assuming that the process of transcription requires a time that can be represented by a random variable following a negative exponential approximation may clearly be an approximation. Still, there are many modeling studies that make this assumption without questioning its accuracy.

The correctness of the assumption can be tested with mathematical arguments, as it has been done for the Michaelis-Menten kinetics, or with wet-lab experiments, which can estimate properties of the stochastic process that represents the production of the transcripts molecules. Modeling studies [10] indicate that the process of transcript production exhibits less variability than a simple Poisson process, the one that one is entitled to

assume if the transcription times followed a negative exponential distribution. However, [10] also shows that when the pausing that occurs in gene transcription is frequent enough, the transcript production process tends to become a Poisson process.

Once more, we remark that, once the fundamental hypothesis is assumed to hold for the reaction $G + P \rightarrow G + T + P$, the SSA algorithm can be applied, and that the quality of the results it provides are only determined by the accuracy of the approximation entailed by the fundamental hypothesis. In the following section we discuss about the expected effects of the approximations introduced when applying the SSA.

5 Impact of approximations

Postulating the validity of the fundamental hypothesis 1) when the same is not applicable can lead to an approximation whose effects are hardly predictable. Indeed, assuming that a random variable follows the negative exponential distribution implies a precise choice of the variance for the process it models. Specifically, if λ is the parameter of a negative exponential distribution, this means that the average value is λ^{-1} and the variance λ^{-2}, that is the only one parameter of the distribution determines both, and no choice is left on the amount of variability to be modeled.

A commonly used measure of the amount of variability of a random variable is the *coefficient of variation*, defined as the ratio between standard deviation (the square root of the variance) and expected value of the variable. The coefficient of variation of a negative exponential random variable is exactly 1. This implies a certain amount of stochastic fluctuation around the average value for a set of values sampled from the distribution. In a biological system, stochastic fluctuations plays an important role, as they determine the probability with which different portions of the state space of the system are accessible from a given state, and ultimately the reachability of certain equilibrium conditions or limit cycles. Thus, a modeling choice based on Gillespie's fundamental hypothesis may result in a simulated dynamics that does not match the real one of the system being modeled.

Examples of such discrepancies are occasionally found in the literature. For instance, the paper [1] compares the results obtained through the SSA for various abstractions of the same biochemical system, clearly showing that assuming the fundamental hypothesis is valid for all of them leads to inconsistent predictions.

A slightly different though related perspective on the matter is found

in [9], where the authors show that a deterministic characterization of a set of reactions in terms of ODEs leads to wrong results, whereas the SSA algorithm can better predict the dynamic evolution of the biochemical system. This paper is interesting because it demonstrates the importance of selecting the proper amount of stochastic fluctuation for properly describing the dynamics of a system. Specifically, [9] shows that the total absence of stochastic fluctuations turns out in a wrong simulated evolution as the model cannot reach the part of the state space that is eventually occupied by the modeled system.

To the best of our knowledge, there is not a general theory to predict a priori the effect that different choices of random variable distributions have on stochastic model results. As a matter of fact, when there is no information on whether the fundamental hypothesis is satisfied or not for a set of reactions, validation of model results is the only means available to determine a posteriori the quality of a model. The results obtained through the SSA, which are exact for the input model, must be compared with the observed behavior of the real system. A positive validation indicates that the modeling hypothesis made on the structure and on the stochastic characterization of the system are indeed valid.

6 Conclusions

This report discusses the exactness of the Stochastic Simulation Algorithm proposed by Gillespie for the simulation of biochemical systems composed on a set of coupled reactions. It reviews the hypothesis under which the exactness of the simulation results has been demonstrated, and points out where approximations and inaccuracies can be introduced.

The key point that determines the accuracy of SSA results is found in the adequacy of the fundamental hypothesis of Gillespie for the system to be studied, which implies that the stochastic process considered to represent the dynamic evolution of the system is a Continuous-Time Markov Chain. Indeed, depending on the specific phenomena considered and on the level of abstraction at which the system is to be modeled, assuming that all reaction times are random variables distributed according to the negative exponential law may be inadequate and introduce approximations.

On the other hand, the SSA is always exact with respect to the CTMC model that is provided to it as an input, as it does not introduce any more hypothesis on the system. Whatever discussion about the prerequisites of applicability of Gillespie's methods is to be moved outside the algorithm

itself, and placed at the level of the stochastic characterization of reaction times. When the available knowledge does not allow a precise characterization of the reaction times, a first cut modeling choice that matches the average values, such as the one based on the negative exponential distributions, may be a valid modeling option. This same choice can be made because of the computational advantages offered by the application of the SSA. At any rate, it is the responsibility of the modeler to make it clear when this choice is an approximation, and to conduct a careful validation of model results to ascertain its adequacy in capturing the interesting behaviors of the modeled system.

References

[1] R. Bundschuh, F. Hayot, and C. Jayaprakash. Fluctuations and slow variables in genetic networks. *Biophysical Journal*, 84(3):1606–1615, 2003.

[2] Y. Cao, H. Li, and L. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting system. *Journal of Chemical Physics*, 121(9):4059–4067, 2004.

[3] J. L. Doob. *Stochastic processes*. John Wiley and Sons, New York, 1953.

[4] M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry*, 104(9):1876–1889, 2000.

[5] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

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

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

[8] C. V. Rao and A. P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. *Journal of Chemical Physics*, 118(11):4999–5010, 2003.

[9] M. S. Samoilov and A. P. Arkin. Deviant effects in molecular reaction pathways. *Nature Biotechnology*, 24(10):1235–1240, 2006.

[10] M. Voliotis, N. Cohen, C. Molina-Paris, and T. Liverpool. Fluctuations, pauses and backtracking in DNA transcription. *Biophysical Journal*, 94(2):334–348, 2007.