STOCHASTIC SIMULATION AND PARAMETER ESTIMATION OF FIRST ORDER CHEMICAL REACTIONS

Katrien De Cock, Xueying Zhang, Mónica F. Bugallo, Petar M. Djurić

Stony Brook University
COSINE
Department of Electrical & Computer Engineering
Stony Brook, New York 11794-2350
phone: +1 631 632 8424, fax: +1 631 632 8494, email: decock@ece.sunysb.edu
web: http://www.ece.sunysb.edu/∼cspv/CSPV.html

ABSTRACT

In this paper, we present fast stochastic simulation methods for the class of first order chemical reactions. The methods are based on the exact distributions for the number of molecules or their Gaussian approximations. Furthermore, using the adopted models, we develop parameter estimation methods for the reaction rates. Although we only discuss two basic reactions, the single channel and reversible first order reactions, the obtained results can be applied to more complex cases.

1. INTRODUCTION

The traditional macroscopic approach to kinetics analysis of enzymatically catalyzed reactions, based on the seminal work of Michaelis and Menten [13], assumes uniform and homogeneous concentrations of reactants and products [3]. Biological systems, such as a cell, however, have low numbers of certain molecules. Small fluctuations become here important [2, 11, 14]. Recent research on reactions inside the cell has relinquished the macroscopic approach and started to model the reactions in a mesoscopic, stochastic way [1]. Indeed, chemical reactions occur as discrete events as a result of molecular collisions that cannot be predicted with certainty. These events are modeled by stochastic processes with the objective to predict the time evolution of the reactions of interest.

The motivation for our research lies in the simulation and parameter estimation of enzyme reactions in lung tissues. It is becoming increasingly evident [9] that macroscopic, deterministic methods fail to explain proteinase-catalyzed destructions of connective lung tissue by inflammatory cells. The actual molecular events are better described by stochastic mathematical models.

There are two main approaches for the stochastic study of chemical reactions: the first obtains the distribution function of the number of molecules using the master equation [8, 12, 17] and the second relies on Monte Carlo simulation methods [4, 10, 16].

In 1976, Gillespie published his stochastic simulation algorithm (SSA), which is a Monte Carlo method for numerically computing the time evolution of a chemical system [4]. The fundamental distribution function for SSA for the time interval between two reactions is a time varying exponential distribution, whose parameter depends on the current number of reactant molecules in the system and the reaction rate parameter. We use SSA in Section 3.3 for the simulation of

data to test our estimator for the reversible reaction rate parameters.

The simulation of reactions by SSA can become very slow since every single event is simulated. This is especially the case when the number of molecules in the system is large, and there are several reactions that can occur at any time. Therefore, Gillespie introduced a faster, but approximate method, the τ-leap method, where bigger time leaps are taken so that more than one reaction takes place per simulation step [5, 6, 15]. However, the time steps should still be rather small for a good approximation.

In this paper we discuss simulation methods that use big time steps and are still very accurate. This is possible because we stay closer to the true distributions than in the τ-leap method. We only deal with two elementary reactions: the single channel first order reaction and the reversible first order reaction. However, our results are applicable to more complex reactions. Often, reactions or a set of chemical reactions that seem to be complex at first sight, can be simplified. A second order reaction where one of the reactants is much more abundant than the other, for example, can be approximated by a first order reaction. Or, if one reaction is slow compared to other reactions, only the slow one is important. (See [18] for an enzyme-substrate reaction where these simplifications were applied successfully.) Furthermore, we first want to study the stochastic properties of basic reactions so that we can combine them later.

The structure of the paper is as follows. In Section 2 we discuss the single channel first order reaction and in Section 3 the reversible first order reaction. For both cases, the distribution of the number of molecules is given, a simulation procedure based on the distribution or its Gaussian approximation is discussed and we show how the reaction parameters can be estimated from measurements. Finally, in Section 4 we give some concluding remarks.

2. THE SINGLE CHANNEL FIRST ORDER REACTION

In this section, we study the single channel first order reaction in a closed system. It is usually represented by the following expression: $A \xrightarrow{k} B$, where k is the reaction parameter.

2.1 Distribution function

Let $X_A(t)$ be the number of A molecules at time t and the initial number of A molecules be denoted by $a_0 = X_A(0)$. Then, the number of reactions happening in this closed system in

the interval $[0,t]$, denoted by $M(t)$, is equal to $a_0 - X_A(t)$. Since $M(t)$ is a pure birth process, one can deduce that the number of reactions in $[0,t]$ has a binomial distribution $(a_0, 1 - e^{-kt})$ (this can also be obtained via the master equation):

$$p_m(t) = \binom{a_0}{m} e^{-(a_0-m)kt}(1 - e^{-kt})^m , \qquad (1)$$

where $t \geq 0$ and $m = 0, \ldots, a_0$. As is well-known (see e.g. [7]), for large a_0, the binomial distribution can be approximated by a discretized normal distribution with the same mean and variance as the binomial distribution, which is very useful in the simulation of the reaction (see further). The mean and variance for $M(t)$ are given as $E(M(t)) = a_0(1 - e^{-kt})$ and $Var(M(t)) = a_0 e^{-kt}(1 - e^{-kt})$, respectively. The Gaussian approximation for (1) is then equal to

$$p_m(t) \approx \frac{1}{D} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(m-\mu)^2}{2\sigma^2}} \sum_{i=0}^{a_0} \delta(m-i) , \qquad (2)$$

where

$$\begin{cases} D = \sum_{i=0}^{a_0} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(i-\mu)^2}{2\sigma^2}} , \\ \mu = a_0(1 - e^{-kt}) , \\ \sigma^2 = a_0 e^{-kt}(1 - e^{-kt}) . \end{cases} \qquad (3)$$

2.2 Simulation

To generate a series of samples $X_A(t_i)$ and $X_B(t_i)$ at time $t_0, \ldots, t_N$ with initial number of A and B molecules equal to a_0 and b_0, respectively, we use the following procedure:

1. $X_A(t_0) = a_0$
2. for $i = 1$ to N
   - substitute a_0 in (2) and (3) (or in (1) for an exact simulation) with $X_A(t_{i-1})$
   - draw a sample m from (2) (or from (1) for an exact simulation) with $t = t_i - t_{i-1}$
   - $X_A(t_i) = X_A(t_{i-1}) - m$.

The number of B molecules at $t_0, \ldots, t_N$ is equal to $X_B(t_i) = a_0 + b_0 - X_A(t_i)$ for $i = 0, \ldots, N$.

Drawing samples directly from the probability mass function (1) constitutes an exact, but possibly very slow method (when the number of molecules is large), while using the truncated Gaussian distribution produces a fast approximation.

2.3 Parameter estimation

Assume that $N + 1$ measurements $X_A(t_i)$ of the number of reactant molecules A at equidistant time points, t_0, t_1, $\ldots$, t_N, where $t_{i+1} - t_i = \tau$, are given. For simplicity, we assume that $X_B(t_0) = 0$. We want to estimate the reaction parameter k from these measurements. Based on the exact distribution of $M(t)$ given in (1), the likelihood function of the unknown parameter k can be written as

$$p(X_A(t_0), \ldots, X_A(t_N); k) = \prod_{i=0}^{N-1} \binom{X_A(t_i)}{m_i} e^{-(X_A(t_i)-m_i)k\tau}(1 - e^{-k\tau})^{m_i} ,$$

where m_i is the number of product molecules produced in $[t_i, t_{i+1}]$. They can be computed from the data as $m_i = X_A(t_i) - X_A(t_{i+1})$.

The maximum likelihood estimator (MLE) of k in terms of the measurements $X_A(t_0), \ldots, X_A(t_N)$ is then given as

$$\hat{k} = -\frac{1}{\tau} \ln\left(1 - \frac{X_A(t_0) - X_A(t_N)}{\sum_{i=0}^{N-1} X_A(t_i)}\right) , \qquad (4)$$

and the Cramér-Rao lower bound (CRLB) for the variance of unbiased estimators of k is equal to

$$\text{CRLB} = \frac{(1 - e^{-k\tau})^2}{\tau^2 e^{-k\tau} X_A(t_0)(1 - e^{-kN\tau})} .$$

The estimator in (4) was applied to an enzyme-substrate reaction in [18]. It was compared to estimators based on a Poisson distribution approximation. More information on its statistical properties (bias and variance) can also be found in [18].

3. THE REVERSIBLE FIRST ORDER REACTION

The chemical notation for a reversible first order reaction is given by $A \underset{k_2}{\overset{k_1}{\rightleftarrows}} B$. The reaction consists of two coupled single channel first order reactions, and the order in which these two reactions take place is random.

3.1 Distribution function

Let $X_A(t)$ be the number of A molecules at time t, x_0 the total number of A and B molecules in the closed system, and k_1 and k_2 the forward and backward rate constants, respectively. Using the master equation approach, we derived the generation function as

$$F(s,t) = \left(\frac{-\lambda(1-s)e^{-kt} + \lambda + s}{(1-s)e^{-kt} + \lambda + s}\right)^{X_A(0)} \times \left(\frac{(1-s)e^{-kt} + \lambda + s}{1 + \lambda}\right)^{x_0} ,$$

where $k = k_1 + k_2$ and $\lambda = \frac{k_1}{k_2}$. The generation function is defined as

$$F(s,t) = \sum_{x=0}^{x_0} p_x(t)s^x ,$$

where $p_x(t)$ $(x = 0, 1, \ldots, x_0)$ is the distribution function for the number of A molecules at time t. Consequently, the values $p_x(t)$ can be obtained by computing the coefficient of the term of order x in the Taylor series of $F(s,t)$ about $s = 0$. For the special case where $X_A(0) = x_0$ and hence $X_B(0) = 0$, the distribution function can be expressed analytically:

$$p_x(t) = \left(\frac{k_1}{k}\right)^{x_0} \binom{x_0}{x} \left(e^{-kt} + \frac{1}{\lambda}\right)^x (1 - e^{-kt})^{x_0-x} .$$

For the general case, an analytic expression for the expected value of the number of A molecules and the variance can be derived from the generation function:

$$E(X_A(t)) = \left(\frac{\partial F}{\partial s}\right)_{s=1} = \frac{(1 - e^{-kt})x_0}{\lambda + 1} + X_A(0)e^{-kt} \qquad (5)$$
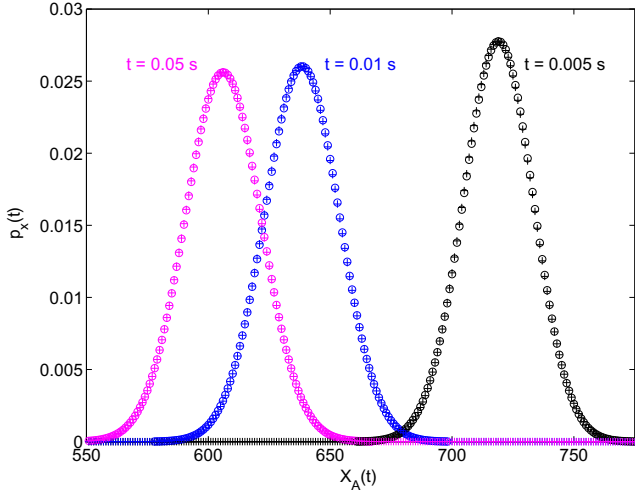
Figure 1: The distribution function for the number of *A* molecules in the first order reversible reaction after $t = 0.005$ s, $t = 0.01$ s and $t = 0.05$ s. The initial values are $X_A(0) = 10^3$ and $X_B(0) = 10$ and the rate parameters are equal to $k_1 = 100$ s^{-1} and $k_2 = 150$ s^{-1}. The circles represent the true distributions and the plus signs the Gaussian approximations.

and

$$Var(X_A(t)) = \left(\frac{\partial^2 F}{\partial s^2}\right)_{s=1} + \left(\frac{\partial F}{\partial s}\right)_{s=1} - \left(\frac{\partial F}{\partial s}\right)_{s=1}^2$$
$$= \frac{1}{(1+\lambda)^2}(1 - e^{-kt})$$
$$\times \left(x_0\lambda + e^{-kt}\left(X_A(0)(\lambda^2 - 1) + x_0\right)\right). \quad (6)$$

A different derivation of the expressions (5) and (6) can be found in the appendix of [15].

Again, the true distribution for the number of *A* molecules can be approximated by the discretized normal distribution with mean and variance equal to the values in (5) and (6):

$$p_x(t) \approx \frac{1}{D}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\sum_{i=0}^{x_0}\delta(x-i), \quad (7)$$

where $D = \sum_{i=0}^{x_0}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(i-\mu)^2}{2\sigma^2}}$, $\mu = E(X_A(t))$ and $\sigma^2 = Var(X_A(t))$.

Example 1 *Assume that $X_A(0) = 10^3$, $X_B(0) = 10$, $k_1 = 100$ s^{-1} and $k_2 = 150$ s^{-1}. We show the true distributions (circles) for $t = 0.005$ s, $t = 0.01$ s and $t = 0.05$ s and their Gaussian approximations (plus-signs) in Figure 1. The normal distributions are clearly very good approximations.*

3.2 Simulation

The simulation procedure for the reversible reaction is similar to that of the single channel reaction in Section 2.2. Assume that the initial number of *A* molecules and *B* molecules are known and equal to a_0 and b_0, respectively. We generate a realization of the evolution of the number of *A* molecules at given time points $t_0, \ldots, t_N$ as follows.
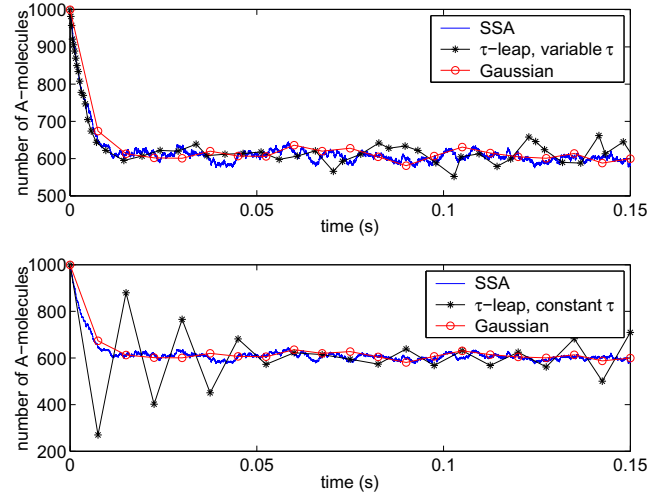

Figure 2: Simulation results for the reversible first order reaction, where $X_A(0) = 10^3$, $X_B(0) = 10$, $k_1 = 100$ s^{-1} and $k_2 = 150$ s^{-1}. In the top picture we show in full line a realization of the stochastic simulation algorithm (SSA) [4], the asterisks represent a realization of the τ-leap method with variable τ [6] and a realization obtained with the simulation method described in Section 3.2 is given by the circles. The only difference in the bottom picture is the τ-leap realization (asterisks). It is obtained with the same constant $\tau = 0.0075$ s as in our Gaussian simulation method.

1. $X_A(t_0) = a_0$ and $x_0 = a_0 + b_0$
2. for $i = 1$ to N
   - substitute $X_A(0)$ in (5) and (6) with $X_A(t_{i-1})$
   - draw a sample x from (7) with $t = t_i - t_{i-1}$
   - $X_A(t_i) = x$

The number of *B* molecules at $t_0, \ldots, t_N$ is equal to $X_B(t_i) = a_0 + b_0 - X_A(t_i)$ for $i = 0, \ldots, N$.

Example 2 *We use the same initial values and rate parameters as in Example 1. The first order reversible reaction is simulated on the time interval $[0, 0.15]$ s in four different ways:*

1. *using the exact stochastic simulation algorithm (SSA) described in [4] (full line in Figure 2),*
2. *using the improved τ-leap method of reference [6] (asterisks in the top picture of Figure 2),*
3. *using the method presented in this section, based on the Gaussian approximation of the distribution functions (circles in Figure 2), where the constant time steps $\tau = t_i - t_{i-1}$ are equal to 0.0075 s,*
4. *using the τ-leap method with constant leap-size (as in [15]) $\tau = 0.0075$ s (asterisks in the bottom picture of Figure 2).*

As can be seen in the top picture of Figure 2, the τ-leap method with variable τ takes more simulation steps than the method presented in this paper (50 vs. 20). Especially in the beginning of the reaction, where the number of A-molecules decreases fast, small simulation steps are taken. In the bottom picture, we show what happens if the same time points are taken in the τ-leap method as in our Gaussian method. The first point at 0.0075 s obtained by τ leap, is much too

τ	N	k_1	k_2	mean $\pm 2\sigma$	
0.1	50	0.5	1	0.51 ± 0.07	1.01 ± 0.10
		1	1	1.00 ± 0.15	1.00 ± 0.12
		1.5	1	1.52 ± 0.28	1.01 ± 0.16
		2	1	2.04 ± 0.34	1.02 ± 0.16
0.025	200	0.5	1	0.50 ± 0.07	1.01 ± 0.10
		1	1	1.00 ± 0.14	1.00 ± 0.12
		1.5	1	1.50 ± 0.19	1.00 ± 0.12
		2	1	2.01 ± 0.27	1.00 ± 0.12

Table 1: The estimation results for different values of the reaction parameters k_1 and k_2.

small. Increasing the time step τ to 0.01 s even leads to negative numbers of molecules, whereas our Gaussian method still performs very well.

3.3 Parameter estimation

Given the number of A molecules at $t_0, \ldots, t_N$, denoted by $X_A(t_i)$ $(i = 0, \ldots, N)$, and the total amount of molecules in the closed system x_0, we want to estimate the reaction parameters k_1 and k_2. Based on the Gaussian approximation for the distribution of A molecules in (7), the likelihood function of the parameters k_1 and k_2 can be written as

$$p(X_A(t_0), \ldots, X_A(t_N); k_1, k_2) = \prod_{i=1}^{N} p(X_A(t_i)|X_A(t_{i-1}); k_1, k_2) ,$$
(8)

where $p(X_A(t_i)|X_A(t_{i-1}); k_1, k_2)$ is the Gaussian distribution in (7) with $D = D_i$ is the normalization factor, the mean is equal to $\mu_i = \frac{(1-e^{-k(t_i - t_{i-1})})x_0}{\lambda+1} + X_A(t_{i-1})e^{-k(t_i - t_{i-1})}$ and the variance is

$$\sigma_i^2 = \frac{1}{(1+\lambda)^2}(1 - e^{-k(t_i - t_{i-1})})$$
$$\times \left(x_0\lambda + e^{-k(t_i - t_{i-1})}(X_A(t_{i-1})(\lambda^2 - 1) + x_0) \right) .$$

The MLE for the parameters k_1 and k_2 can then be obtained as $(\hat{k}_1, \hat{k}_2) = \underset{k_1, k_2}{\arg\min} \left\{ \sum_{i=1}^{N} 2\ln D_i + \ln \sigma_i^2 + \frac{(X_A(t_i) - \mu_i)^2}{\sigma_i^2} \right\}$. Since it is not possible to find an analytical expression for the maximum of the likelihood function (8), we used a numerical optimization method to estimate the parameters.

Example 3 *Estimation results of the reaction parameters k_1 and k_2 are given in Table 1. The parameter τ denotes the time interval between two samples. The parameter N denotes the number of samples taken in each realization. The number of realizations used for the estimation of the mean and variance is 100. The simulated data used for estimation are generated with the SSA [4]. The initial condition for the simulation is: $X_A(0) = 10^3$ and $X_B(0) = 0$.*

4. CONCLUSIONS AND FUTURE WORK

In this paper we discussed the stochastic properties of two basic reactions, the single-channel and reversible first order reactions. Based on their distribution functions, we proposed fast and accurate stochastic simulation methods and showed how the reaction parameters can be estimated from measured data points.

Currently, we are extending this study to second order reactions. The true challenge, however, will be to tackle sets of coupled chemical reactions. Finally, we will apply our results to stochastic modeling, simulation and parameter estimation of enzyme reactions in lung tissues.

REFERENCES

[1] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage λ−infected *escherichia coli* cells", *Genetics*, vol. 149, pp. 1633–1648, 1998.

[2] W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins, "Noise in eukaryotic gene expression", *Nature*, vol. 422, pp. 633–637, 2003.

[3] A. Cornish-Bowden, *Fundamentals of Enzyme Kinetics*. Portland Press, 1995.

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

[5] D. T. Gillespie, "Approximate accelerated stochastic simulation of chemically reacting systems", *Journal of Chemical Physics*, vol. 115, pp. 1716–1733, 2001.

[6] D. T. Gillespie and L. R. Petzold, "Improved leap-size selection for accelerated stochastic simulation", *Journal of Chemical Physics*, vol. 119, pp. 8229–8234, 2003.

[7] P. G. Hoel, *Introduction to Mathematical Statistics*. John Wiley & Sons, 5th edition, 1984.

[8] I. J. Laurenzi, "An analytical solution of the stochastic master equation for reversible bimolecular reaction kinetics", *Journal of Chemical Physics*, vol. 113, pp. 3315–3322, 2000.

[9] T. G. Liou and E. J. Campbell, "Nonisotropic enzyme–inhibitor interactions, a novel nonoxidative mechanism for quantum proteolysis by human neutrophils", *Biochemistry*, vol. 34, pp. 16171–16177, 1995.

[10] J. J. Lukkien, J. P. L. Segers, and P. A. J. Hilbers, "Efficient Monte Carlo methods for the simulation of catalytic surface reactions", *Physical Review E*, vol. 58, pp. 2598–2610, 1998.

[11] H. H. McAdams and A. Arkin, "It's a noisy business", *Trends in Genetics*, vol. 15, pp. 65–69, 1999.

[12] D. A. McQuarrie, "Stochastic approach to chemical kinetics", *Journal of Applied Probability*, vol. 4, pp. 413–478, 1967.

[13] L. Michaelis and M. L. Menten, "Die kinetik der invertinwerkung", *Biochemische Zeitschrift*, vol. 49, pp. 333–369, 1913.

[14] C. V. Rao, D. M. Wolf, and A. P. Arkin, "Control, exploitation and tolerance of intracellular noise", *Nature*, vol. 420, pp. 231–237, 2002.

[15] M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie, "Stiffness in stochastic chemically reacting systems, the implicit tau-leaping method", *Journal of Chemical Physics*, vol. 119, pp. 12784–12794, 2003.

[16] A. B. Stundzia and C. J. Lumsden, "Stochastic simulation of coupled reaction-diffusion processes", *Journal of Computational Physics*, vol. 127, pp. 196–207, 1996.

[17] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*. North-Holland, 2001.

[18] X. Zhang, K. De Cock, M. F. Bugallo, and P. M. Djuric, "Stochastic simulation and parameter estimation of enzyme reaction models", in *Proceedings of the 2003 IEEE Workshop on Statistical Signal Processing*, St. Louis, MO, 2003.