STOCHASTIC SIMULATION AND PARAMETER ESTIMATION OF ENZYME REACTION MODELS

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

Department of Electrical & Computer Engineering
Stony Brook University
Stony Brook, NY 11794-2350
E-mail: {sherry,decock,monica,djuric}@ece.sunysb.edu

ABSTRACT

The development of models and estimators that can satisfactorily describe enzyme reactions are extremely valuable to understand life processes. In this paper, we address the problem of modeling and parameter estimation of enzyme reactions from a stochastic perspective. The simulation results show that obtained estimators are adequate and accurate enough for this type of systems.

1. INTRODUCTION

Enzymes are the biological catalysts responsible for supporting almost all the chemical reactions that maintain the human body in a regular order. For instance, when substances like bacteria, viruses, dust and smoke enter the lungs, white blood cells containing the enzyme elastase migrate to the site of infection helping the digestion of the invaders. Due to their important role in maintaining life processes, the development of adequate dynamic models that describe this kind of systems is critical.

Classical methods for the analysis of enzyme kinetics [1] are constrained to well-stirred systems, which is clearly not the case in the cell mediated processes [2]. In this paper we model experimental data using probabilistic algorithms [3, 4]. Using as starting point the model proposed by Gillespie [3], that has been used in numerous studies [5, 6, 7, 8], we propose two alternative models that are adequate for estimation of parameters.

The main goal is therefore to predict and estimate the unknowns of interest, meaning the reaction constants and the amount of molecules of the reactants, that are used to formulate the mathematical description of the reactions. Parameter estimation using the stochastic model [3] has, to our knowledge, only been tackled by Gibson [9]. However, in his work the complete trajectories of the amount of molecules are considered to be known, which is a very stringent assumption in practice. In our paper, only a limited number of discrete-time measurements is assumed to be available. Computer simulations show that the obtained estimators are adequate for the considered system.

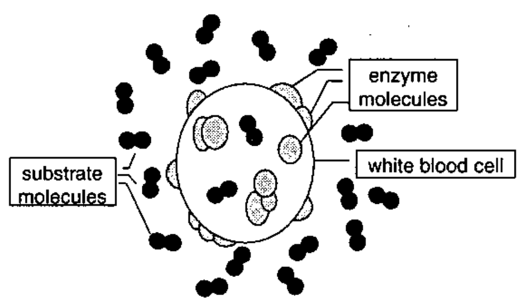

Fig. 1. One cell (e.g. a white blood cell) on which enzyme molecules are immobilized. The substrate molecules float around freely.

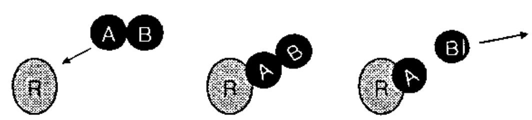

Fig. 2. Illustration of the chemical reaction. The substrate molecule AB approaches the enzyme molecule R (left). The enzyme recognizes the amino acid sequence of A (middle). The enzyme destroys the bond between A and B, A remains attached to R and B is released into the solution (right).

2. PROBLEM DESCRIPTION

Let us consider an enzyme reaction where a soluble substrate with an A-B structure[1] reacts with immobilized

[1]The substrate molecule is composed of two parts: A that will remain attached to the enzyme after the reaction and B that will be released into the solution and will constitute the final product.

0-7803-7997-7/03/$17.00 © 2003 IEEE

enzyme molecules located on the surface of cells (see Figure 1). The formulation of this chemical reaction is given by

$$AB + R \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ARB \xrightarrow{k_{cat}} AR + B\,, \qquad (1)$$

where AB is the substrate molecule, R is the enzyme molecule and ARB is the enzyme-substrate complex, known as the intermediate product. This reaction is illustrated in Figure 2. The objective is to accurately simulate the considered reaction and to estimate the unknowns of interest, meaning the reaction constants and the amount of molecules of the reactants, using the available data given by discrete measurements of B molecules.

3. MATHEMATICAL FORMULATION

A very accurate simulation of the chemical reaction in (1) is obtained by the stochastic algorithm proposed by Gillespie [3]. We will refer to this simulation model as ***model A***.

In the case we focus on, almost all free enzyme R transforms into the ARB complex in a negligibly short period of time. Therefore, the mathematical description that we use for estimation can be simplified to

$$ARB \xrightarrow{k_{cat}} AR + B \qquad (2)$$

where the initial number of ARB molecules in (2) is considered to be equal to the initial number of R molecules in (1). For this simplified description, the aim is to estimate the reaction parameter k_{cat} from the available measurements of X_B.

Reaction (2) can be analyzed by using two models, which we will refer to as ***model*** $\mathbf{B_1}$ and ***model*** $\mathbf{B_2}$. Let the number of times that reaction (2) occurs in the time interval $[t_i, t_{i+1}]$ be M_i. Then, we represent, as in [4], the distribution of M_i by a Poisson probability mass function with mean (and variance) equal to $\lambda(t_i)(t_{i+1} - t_i) = \lambda(t_i)\Delta t_i$, where $\lambda(t) = k_{cat}X_{ARB}(t)$, $X_{ARB}(t)$ is the number of complex molecules at time t, and Δt_i is the sample time interval. We refer to this representation as to ***model*** $\mathbf{B_1}$. Note that we can also take $\lambda(t_i) = k_{cat}X_{ARB}(t_{i+1})$ as the Poisson rate parameter, where the number of complex molecules at the end of each time interval is used. This modified model will be referred to as the 'backward' version of *model* B_1 in analogy to the forward and backward Euler method for the discretization of differential equations.

However, since $X_{ARB}(t)$ is a simple death process, the exact distribution of M_i can be derived. The probability that m reactions occur in the time interval $[t_i, t_{i+1}]$ is:

$$P(M_i = m) = \binom{x_{0i}}{m} e^{-(x_{0i}-m)k_{cat}\Delta t_i}(1 - e^{-k_{cat}\Delta t_i})^m\,, \qquad (3)$$

where $x_{0i} = X_{ARB}(t_i)$. The model based on (3) will be referred to as ***model*** $\mathbf{B_2}$.

4. PARAMETER ESTIMATION

4.1. Estimation of the reaction rate k_{cat}

Assume that $N+1$ measurements of the number of B molecules, denoted by X_B, at time instants $t_0, \ldots, t_N$ and the initial number of enzyme molecules, $X_R(t_0)$, are available and that there are no B molecules at t_0. The maximum likelihood estimate (MLE) of k_{cat}, based on *model* B_1, is then equal to

$$\hat{k}_{cat} = \frac{X_B(t_N) - X_B(t_0)}{\sum_{i=0}^{N-1}(X_R(t_0) - X_B(t_i))\Delta t_i}\,, \qquad (4)$$

and will be called the ***forward estimator***. The Cramér-Rao lower bound (CRLB) for the variance of $\hat{k}_{cat}$ is given by

$$Var(\hat{k}_{cat}) \geq \frac{k_{cat}^2}{X_R(t_0)\left(1 - \prod_{i=0}^{N-1}(1 - k_{cat}\Delta t_i)\right)}\,. \qquad (5)$$

Note that the MLE of k_{cat} based on the backward *model* B_1 (called ***backward estimator***) is equal to that of (4) substituting $X_B(t_i)$ by $X_B(t_{i+1})$. However, the CRLB for this model is also given by (5).

When all the measurements are made uniformly in time, i.e. the time interval between two measurements is $\Delta t_i = \tau$, $i = 0, \ldots, N-1$, then the MLE of k_{cat}, based on *model* B_2, has an analytic solution which is

$$\hat{k}_{cat} = -\frac{1}{\tau}\log\left(1 - \frac{X_B(t_N) - X_B(t_0)}{\sum_{i=0}^{N-1} X_R(t_0) - X_B(t_i)}\right)\,, \qquad (6)$$

and the CLRB is equal to:

$$\text{CRLB} = \frac{(1 - e^{-k_{cat}\tau})^2}{\tau^2 e^{-k_{cat}\tau} X_R(t_0)(1 - e^{-k_{cat}N\tau})}\,. \qquad (7)$$

Note that the MLE for B_1 is a first order approximation of the MLE for B_2[2], which is good for small sampling time intervals.

[2] The first order approximation is given by $\log\frac{1}{1-x} = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \cdots$

4.2. Estimation of both k_{cat} and $X_R(t_0)$

For immobilized enzymes, it is usually very difficult to get an accurate measurement of the number of molecules present at time t_0. In such case, the *forward estimator* based on *model* B_1 for $X_R(t_0)$ and k_{cat} is given by

$$\begin{cases} \hat{X}_R(t_0) = \arg\min_{X_R(t_0)} \left\{ \left| \sum_{i=0}^{N-1} \frac{(X_B(t_{i+1}) - X_B(t_i))}{X_R(t_0) - X_B(t_i)} \right. \right. \\ \qquad \left. \left. -(t_N - t_0) \cdot \frac{X_B(t_N) - X_B(t_0)}{\sum_{i=0}^{N-1} (X_R(t_0) - X_B(t_i)) \Delta t_i} \right| \right\} \\ \hat{k}_{cat} = \frac{X_B(t_N) - X_B(t_0)}{\sum_{i=0}^{N-1} (\hat{X}_R(t_0) - X_B(t_i)) \Delta t_i} . \end{cases} \tag{8}$$

The *backward estimator* has the same form as (8) except that all $X_B(t_i)$ in the denominators must be changed to $X_B(t_{i+1})$.

5. ESTIMATION RESULTS

5.1. Statistical properties of MLEs

To test the statistical properties of the MLEs in (4) and (6), simulation data were generated based on *model* B_1 and *model* B_2, respectively. The initial number of enzyme molecules $X_R(t_0)$ was 10,000 and the parameter k_{cat} was set to different values: 50, 100, 150 and 200. All time intervals Δt_i were given by $\Delta t = \frac{3.6}{N k_{cat}}$, where $N = 100$, which means that each realization consisted of 101 measurements of X_B at time instants $t_0, \ldots, t_{100}$. The number of realizations generated with the same parameters is denoted by I.

The MLEs of k_{cat} of *forward* B_1 in (4) and B_2 in (6) are shown in Table 1 and Figure 3. When the number of realizations I increases, the variance of $\hat{k}_{cat}$ approaches the CRLB. Thus, the *forward estimator* is asymptotically efficient.

Table 2 shows the estimation results of the *forward estimator* in (8) when both k_{cat} and $X_R(t_0)$ are estimated. When compared with the CRLB, we observe that this estimator performs well.

5.2. Estimation results on data from *model A*

In this section we show the estimation results obtained for *model B*, based on data generated using *model A*. Since *model A* is an accurate simulation of reaction (1), simulation data generated by this model follow true experimental data closely. However, the simulation with this model gives too many data with very small time intervals. A sampling procedure is thus needed to produce the desired number of data with appropriate time intervals to take into account the measurement conditions.

The simulated data for the results in Table 3 and Table 4 are obtained as follows. First, we generated data with *model A*, considering $X_R(0) = 10,000$, the concentration of the substrate equal to 8000 μ M, and $X_{AR}(0) = X_{ARB}(0) = X_B(0) = 0$. The reaction constants were set to be $k_{-1} = 1\ \mathrm{s}^{-1}$ and $k_1 = \frac{k_{cat}+k_{-1}}{40}\ (\mu\mathrm{M})^{-1}\mathrm{s}^{-1}$, and the reaction volume was 0.01 l. The parameter k_{cat} was taken equal to 50, 100, 150 and 200 s^{-1}. Next, the simulated data were sampled from time $t_0 = 0$ with $\Delta t_i = \frac{3.6}{N k_{cat}}$ and $N = 200$. In this way, with different k_{cat}, the mean number of reactions (2) recorded, was still approximately constant.

model	k_{cat}	$I = 100$	$I = 1\,000$	$I = 10\,000$
forward B_1	50	-0.10	-0.035	-0.00090
	100	0.040	0.013	-0.0043
	150	-0.074	-0.015	0.020
	200	-0.15	0.0041	0.016
B_2	50	0.14	0.042	0.047
	100	-0.078	-0.022	0.0038
	150	-0.18	0.018	0.049
	200	-0.091	0.021	0.0075

Table 1. The estimated bias of the MLEs (*forward estimator* B_1 in (4) and *estimator* B_2 in (6)) for different values of the reaction parameter ($k_{cat} = 50, 100, 150, 200$). The initial number of enzyme molecules was equal to $10,000$. The entries are the relative values expressed in %, i.e. bias $= 100\ \mathrm{E}\left[\frac{\hat{k}_{cat} - k_{cat}}{k_{cat}}\right]$, where $\mathrm{E}[\cdot]$ denotes expectation. The parameter I denotes the number of realizations used for the estimation of the bias.

	$\hat{k}_{cat}$		$\hat{X}_R(t_0) \cdot 10^{-3}$	
k_{cat}	mean	var.	mean	var.
50	50.00	0.43	10.00	0.47
100	100.0	1.6	10.00	0.45
150	149.9	3.5	10.00	0.43
200	199.9	7.0	10.00	0.43

Table 2. Estimation results for the *forward estimator* in (8). The estimated mean and variance of the estimates for k_{cat} and $X_R(t_0)$ are shown for different values of the reaction parameter ($k_{cat} = 50, 100, 150, 200$). The initial number of enzyme molecules was equal to $10,000$. For estimation of the mean and variance, $1,000$ different realizations were used.

In Table 3 we give the estimated mean and variance of $\hat{k}_{cat}$ obtained by the B_1 estimator in (4) and by the B_2 estimator in (6). Table 4 shows the mean and variance of $\hat{k}_{cat}$ and $\hat{X}_R(t_0)$ by the backward version of the B_1 estimator in (8). From Table 3 we see that the mean of $\hat{k}_{cat}$ by the *forward estimator* is smaller than the real value of k_{cat}. This is due to the simplification of the stochastic model. The *backward estimator* does not have this bias.

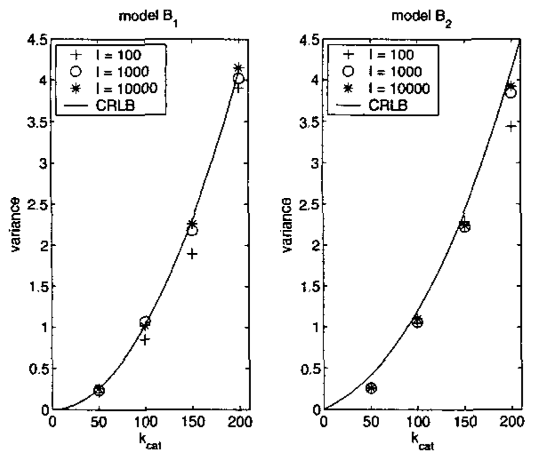

Fig. 3. The estimated variance of the MLE of k_{cat} (left the *forward estimator* B_1 in (4) and right the *estimator* B_2 in (6)) and the CRLBs (full line) for different values of the reaction rate, $k_{cat} = 50, 100, 150, 200$. The initial number of enzyme molecules was equal to $10,000$. The symbols indicate how many realizations were used to estimate the variance ($I = 100$ is denoted by a plus-sign, $I = 1,000$ by a circle and $I = 10,000$ by a star).

6. CONCLUSIONS

In this paper we study models that represent data of enzyme reactions. An accurate stochastic model was used for stochastic simulation of such data. For the estimation of the reaction rate parameter based on discrete-time measurements, the model was simplified to *model* B_1 (forward and backward) and *model* B_2. While in *model* B_2 we used the exact distribution for the number of reactions (2), *model* B_1 approximates it with a Poisson distribution. However, for more complicated reactions, or sets of coupled reactions, it is not possible to derive the exact distribution. The Poisson approximation could still then be used. The simplified models also provide a faster, but less accurate way to simulate the reactions (1) than simulating with *model A*. We noted that for the reactions considered, the backward model based on the Poisson distribution, was superior to the forward model and even to *model* B_2 (see Table 3).

	B_1 *forward*		B_1 *backward*		B_2	
k_{cat}	mean	var.	mean	var.	mean	var.
50	48.36	0.22	49.22	0.28	48.78	0.27
100	96.47	0.96	98.5	1.0	97.63	0.99
150	145.0	1.8	147.6	2.0	146.28	1.93
200	193.5	3.6	196.7	4.3	194.94	4.11

Table 3. Estimation results obtained with the *forward* and *backward* B_1 estimators and with the estimator based on *model* B_2, for data generated using *model A*. The estimated mean and variance of $\hat{k}_{cat}$ are given for different values of the reaction parameter ($k_{cat} = 50, 100, 150, 200$). The initial number of enzyme molecules was equal to $10,000$. For estimation of the mean and the variance, 100 different realizations were used.

	$\hat{k}_{cat}$		$\hat{X}_R(t_0) \cdot 10^{-3}$	
k_{cat}	mean	var.	mean	var.
50	48.59	0.46	10.04	0.56
100	97.2	1.5	10.04	0.58
150	145.9	3.2	10.03	0.44
200	194.5	6.5	10.03	0.49

Table 4. *Backward* estimation results for *model A*. The estimated mean and variance of $\hat{k}_{cat}$ and $\hat{X}_R(t_0)$ are given for different values of the reaction parameter ($k_{cat} = 50, 100, 150, 200$) where the initial number of enzyme molecules was $10,000$. For estimation of the mean and variance, 100 different realizations were used.

7. REFERENCES

[1] L. Michaelis and M. L. Menten, "Die kinetik der invertinwerkung," *Biochemische Zeitschrift*, vol. 49, 1913.

[2] 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.

[3] 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.

[4] D. T. Gillespie, "Approximate accelerated stochastic simulation of chemically reacting systems," *Journal of Chemical Physics*, vol. 115, no. 4, pp. 1716–1733, July 2001.

[5] P. Hannuse and A. Blanche, "A Monte Carlo method for large reaction-diffusion systems," *Journal of Chemical Physics*, vol. 74, pp. 6148–6153, 1981.

[6] H. P. Breuer and F. Petruccione, "How to build master equations for complex systems," *Continuum Mechanics and Thermodynamics*, vol. 7, pp. 439–473, 1995.

[7] M. A. Matias, "On the effects of molecular fluctuations on models of chemical chaos," *Journal of Chemical Physics*, vol. 102, pp. 1597–1606, 1995.

[8] H. H. McAdams and A. Arkin, "Stochastic mechanisms in gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, pp. 814–819, 1997.

[9] M. A. Gibson, *Computational Methods for Stochastic Biological Systems*, Ph.D. thesis, California Institute of Technology, Pasadena CA, 2000.