MAXIMUM LIKELIHOOD BASED METHODS FOR ESTIMATING REACTION RATE CONSTANTS FROM TIME-SERIES

Kenzie D. MacIsaac and Stephen W. Davies
Institute of Biomaterials and Biomedical Engineering, Edward S. Rogers Sr. Dept. of Electrical and Computer Engineering, University of Toronto, CANADA

I. INTRODUCTION

Rational design of cellular control and instrumentation at the molecular level frequently requires stochastic modeling [1]. These simulations require that measurements of the rate constants of the processes involved be available [2]. Unfortunately, evaluation of the chemical reaction rate constants in stochastic biological systems poses certain challenges not present in typical kinetic investigations. Small systems where significant statistical fluctuations occur in the reactant concentration yield time-series concentration data that is noisy. In addition, it is often difficult to examine specific processes in isolation, and as such a single species may undergo several distinct reactions. In such cases it is said that the reactant may pass through several different reaction channels. In this paper the analysis of time-series concentration data is treated probabilistically. This allows for the application of estimation theory to the problem of rate parameter determination in systems where reactant concentration is not a smooth, deterministic function of time.

II. STATISTICAL MODEL

Consider a system of M elementary chemical reactions occurring in a volume V. Defining c_μ as the rate constant for reaction μ, the average probability that a particular set of molecules will react according to reaction μ in the differential time interval dt is $c_\mu dt$. If there are h_μ distinct combinations of reactants that can react according to μ then the average probability of reaction μ occurring in volume V over time interval dt is given by:

$$P_{avg}(\mu) = h_\mu c_\mu dt \qquad (1)$$

A system of reactions is coupled when the completion of one reaction event affects the probability of the other reactions by changing reactant numbers. For a system of coupled chemical reactions we define a *reaction probability density function* $P(\tau,\mu)d\tau$ that gives the probability at time t that the next reaction will occur in the interval $[t+\tau, t+\tau+d\tau]$ and that the reaction will be of type μ. The reaction probability density function can be calculated as the product of $h_\mu c_\mu dt$ and the probability $P_0(\tau)$, that the time interval $[t,t+\tau]$ will be reaction free.

Following Gillespie [2], the time interval τ, may be divided into K small intervals of length $\varepsilon = \tau/K$. The interval is assumed to be small enough so that only one reaction can occur within it. The binomial theorem may then be used to calculate the probability of no reaction occurring in the small interval:

$$\prod_{\mu=1}^{M}[1-h_\mu c_\mu \varepsilon] = 1-\sum_{\mu=1}^{M} h_\mu c_\mu \varepsilon \qquad (2)$$

Since there are K intervals we can write:

$$P_0(\tau) = \left[1-\sum_{\mu=1}^{M} h_\mu c_\mu \frac{\tau}{K}\right]^K \qquad (3)$$

Taking the limit of the above equation as $K \to \infty$ and using the limit formula for the exponential function gives:

$$P_0(\tau) = \exp\left[-\sum_{\mu=1}^{M} h_\mu c_\mu \tau\right] \qquad (4)$$

Now multiplying (1) and (4) yields:

$$P(\tau,\mu) = h_\mu c_\mu \exp\left[-\sum_{\upsilon=1}^{M} h_\upsilon c_\upsilon \tau\right] \qquad (5)$$

So the reaction probability density function consists of a series of discrete exponentials with different weightings that depend on the propensity of each reaction occurring. For a single reaction, (5) reduces to a simple one-dimensional exponential distribution.

III. MAXIMUM-LIKELIHOOD ESTIMATORS FOR COMPLETELY OBSERVED SYSTEMS

A reacting system is completely observed when the creation and decay of individual molecules of reacting species can be monitored and recorded. Consider the irreversible isomerization reaction, with rate parameter c, shown below:

$$Y \xrightarrow{C} Z$$

Reaction 1. A simple isomerization reaction with rate parameter c.

Chemical systems composed entirely of elementary reactions are Markov-1 processes. The reaction times are therefore independent random variables. A series of data points (Y_i, t_i) are collected for reactant Y corresponding to the reaction times (t_i) and number of Y molecules present at those times

(Y_i). The conditional likelihood of this observation is:

$$P(Data / c) = \prod_{i=1}^{N} Y_i c \exp\left(-Y_i c t_i\right) \quad (6)$$

Taking the log of both sides of (6), and maximizing with respect to c yields the maximum likelihood estimate:

$$\widehat{c}_{ML} = \frac{N}{\sum_{i=1}^{N} Y_i t_i} \quad (7)$$

Since the mean of an exponential distribution with decay parameter k is given by 1/k, this result can be interpreted as the inverse of the weighted average of reaction times. Calculating the Cramer-Rao lower bound for the maximum-likelihood estimate of c gives:

$$\mathrm{var}[\widehat{c}_{ML} - c] \geq \frac{c^2}{N} \quad (8)$$

This result suggests that the variance on the estimate of c decreases with the number of sample points, and that the variance increases with the square of the rate parameter c. Better performance from the maximum likelihood estimator is expected for slow reactions rather than for fast reactions.

When a fully observed reactive species is both created and degraded through two individual reaction channels, the result of (7) can be applied by grouping the data into separate production and degradation sets. Consider the set of coupled Lotka reactions commonly used to model predator-prey relationships:

$$Y \xrightarrow{C_1} 2Y$$

$$Y+Z \xrightarrow{C_2} 2Z$$

$$Z \xrightarrow{C_3} \text{Degraded}$$

Reaction 2. The Lotka reactions. Y is a self-replicating species. Z consumes Y to multiply and is itself degraded spontaneously.

If reactant Y is fully observed and we group the N reaction event data into M production and (N-M) degradation events as described above, then the probability of the data is:

$$P(Data / c_1, c_2) = \prod_{i=1}^{M} Y_i c_1 \exp\left((Z_i c_1 + Y_i c_2) t_i\right) \times \prod_{i=M+1}^{N} Z_i Y_i c_2 \exp\left((Z_i c_1 + Y_i c_2) t_i\right) \quad (9)$$

Following the same procedure used to derive (7), it is seen that the maximum likelihood estimate of c_1 is given by:

$$\widehat{c}_{1,ML} = \frac{M}{\sum_{i=1}^{N} Y_i t_i} \quad (10)$$

The Cramer-Rao lower bound for this system is given by:

$$\mathrm{var}[\widehat{c}_{1,ML} - c_1] \geq \frac{c_1^2}{M} \quad (11)$$

The lower bound on the variance of the estimate is dependent on the observed number of reactions of type 1, not the total number of reactions.

A series of 100 simulations of the Lotka reactions was performed with rate parameters c_1=10, c_2=0.01 and c_3=10 and initial conditions Y=1000 and Z=1000. For each simulation, 10000 reaction events were recorded. A representative result of these simulations is shown in the figure below:

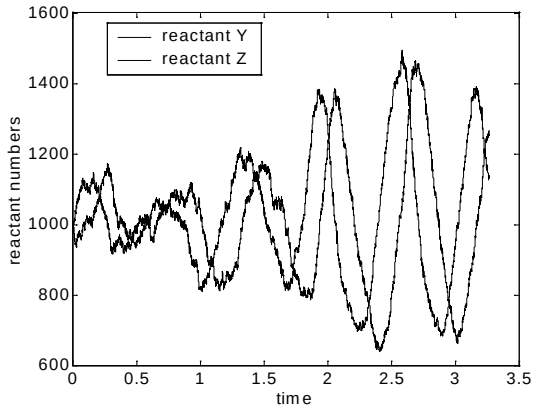

Figure 1. Stochastic simulation of 10000 Lotka reaction events.

As Figure 1 illustrates, reactant levels are inherently noisy for stochastic chemical systems. The maximum-likelihood estimator of (10) was applied to each of the 100 sets of simulation results to obtain estimates of rate parameter c_1. The mean estimate of c_1 was 10.0036. The variance on the estimates was 0.0261. These results are consistent with what would be expected from (10) and (11).

IV. SPECIES REACTING THROUGH MULTIPLE CHANNELS

When a fully observed chemical species is produced or degraded through multiple pathways, the rate parameter estimation problem is complicated by the fact that we cannot know with certainty which reaction has occurred. In this case, rate parameter estimation becomes a hidden data problem and iterative techniques must be used to obtain estimates.

The Expectation Maximization (EM) algorithm is a general approach for maximum likelihood estimation of parameters in statistical models with

hidden variables [3]. In EM, the complete data is viewed as consisting of a set of observable variables x_i, and the hidden variables w_i. The probability model is $p(x,w|\theta)$, where θ is the vector of parameters to estimate. As the vector **w** is not observed, the complete log likelihood is a random quantity and cannot be maximized directly. In EM an averaging distribution, $q(w|x)$, is used to "average out" the randomness. Setting the averaging distribution equal to the probability of the hidden variable conditioned on the observed variables, $P(w|x)$, allows the formulation of an auxiliary function which provides a tight lower bound on the complete log likelihood:

$$\pounds(q,\theta) = \sum_z q(w \mid x) \log \frac{p(x,w \mid \theta)}{q(w \mid x)} \quad (12)$$
$$\leq \log p(x \mid \theta)$$

The EM algorithm has two phases or steps. In the E step, the auxiliary function is maximized with respect to the averaging distribution $q(w|x)$. In the M step, the auxiliary function is maximized with respect to the parameters we are trying to estimate. This is repeated iteratively until convergence.

The EM algorithm can be applied to the problem of rate parameter estimation for chemical species reacting through multiple channels by modeling the observed data as a mixture of exponential. Consider species Y produced through three separate channels with rate parameters c_1, c_2, and c_3 as shown below:

$$Y \xrightarrow{c_1} 2Y$$
$$Z \xrightarrow{c_2} Y$$
$$P \xrightarrow{c_3} Y$$

Reaction 3. Multi-channel reaction pathway. Reactant Y is a self-replicating species. Reactants Z and P are converted to Y through irreversible isomerization reactions.

In the case where Y is fully observed, and the concentrations of reactants Z and P can be measured as a function of time, the probability of the observed data can be modeled as shown below:

$$P(Data / \bar{c}, \bar{\theta}) = \prod_{i=1}^{N} (\theta_1 Y_i c_1 + \theta_2 Z_i c_2 + \theta_3 P_i c_3) \exp\left(-\sum_{\mu=1}^{3} h_{\mu_i} c_\mu t_i \right) \quad (13)$$

The θ's in (13) are the mixture densities of each exponential reaction process. The h_μ's for data point i are given by the number of molecules of Y, Z, and P for μ = 1, 2, and 3 respectively. They are denoted in (13) by Y_i, Z_i, and P_i.

Now, if the observed data **x**, is treated as the reactant concentrations of Y, Z, and P and the times of reaction for Y, and the hidden data **w** indicates the reaction channel, then the application of EM becomes straightforward. The averaging distribution $q^{(i)}{}_\mu(w|x)$ is simply the probability that observation i is a result of reaction process μ, where the $q^{(i)}$'s normalize to 1 when summed over the μ reactions. Derived using the techniques of section III, the EM update equations for the rate parameters at iteration m are:

$$\hat{c}_\mu^{(m)} = \frac{\sum_{i=1}^{N} q_\mu^{i,(m)}(w \mid x)}{\sum_{i=1}^{N} h_\mu^i t^i} \quad (14)$$

The Reaction 2 system was simulated stochastically for 20000 events with rate parameters c_1=0.02, c_2=0.5 and c_3=0.1, and initial conditions Y=10, Z=5000, and P=10000. The results of a typical simulation are shown below:

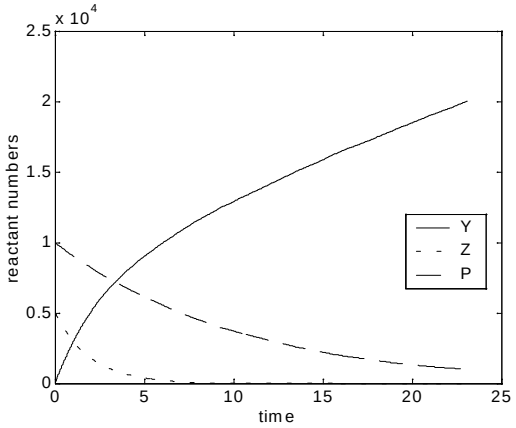

Figure 2. Stochastic simulation of 10000 Reaction 3 events.

The data were analyzed using the EM to extract estimates of the rate parameters. After 50 iterations of EM, the estimates obtained were c_1=0.020, c_2=0.487 and c_3=0.103. This result suggests that EM provides good estimates of the reaction parameters in systems where individual reaction pathways cannot be experimentally distinguished.

V. TIME-SAMPLED SYSTEMS

The results derived in sections III and IV are applicable to chemical systems assuming complete observation of one or more reactants. A more realistic situation is one in which individual reaction times are not observed, but rather the reactant concentrations are sampled at regular time intervals.

In a given time interval T, the probability of observing n events arising from an exponential process with decay parameter k is governed by the Poisson distribution. If the reactant concentrations of a system could be kept constant, then the observed

number of reactions in a time interval T would therefore be Poisson distributed. With a high enough sampling frequency it is reasonable to assume that the total number of reactive species is approximately constant over the sampling interval and that the number of reaction events is approximately Poisson distributed. This assumption also allows the coupled reactions to be treated as statistically independent events.

Unfortunately, for a species reacting through more than one channel, each observed change in reactant number can arise from an infinite number of combinations of reaction events. If the observations are treated as arising from a mixture of Poisson distributions and the hidden data is an indicator function that tells us which mixture component a particular data point arose from, then an EM procedure can be derived to obtain estimates of the rate parameters.

Consider the Lotka reactions shown in Reaction 2. The concentrations of species Y and Z in the reacting system are sampled with a sampling period T. The observed data is defined as the change in the amount of species Y at each time point. Using a Poisson mixture model, an expression for the probability of the observed data is given by:

$$P\left(Data\right) = \prod_i \left[\sum_j \begin{pmatrix} \theta_j \dfrac{\left(h_{i,1}c_1T\right)^{j-1}\left(h_{i,2}c_2T\right)^{m_{i,j}}}{\left(j-1\right)!\left(m_{i,j}\right)!} \times \\ \exp\left(-h_{i,1}c_1T - h_{i,2}c_2T\right) \end{pmatrix} \right] \quad (15)$$

where $m_{i,j}$ is the number of reaction 2 events that must have occurred given the observation and the assumption that j reaction 1 events occurred. The hidden data **w** is a vector of indicator functions denoting the mixture component corresponding to each observation. Applying the averaging distribution, and forming the expected log likelihood of the data allows the EM update equations to be derived:

$$c_\mu^{t+1} = \frac{\sum_i \left(i \sum_j q_{i,j}^t \right)}{T \sum_j h_{\mu,j}} \quad (16)$$

This result was tested by sampling the 100000 event stochastic simulation of the Lotka reactions shown in Figure 1, at a sampling rate of 3.27×10^{-4}s to obtain 10000 data points spaced equally over the total simulation time. The EM algorithm with 20 mixture components was applied to the data. After 300 iterations, the rate parameter estimates converged to c_1=12.76 and c_2=0.013. A plot of the estimates as a function of iteration number is shown below:

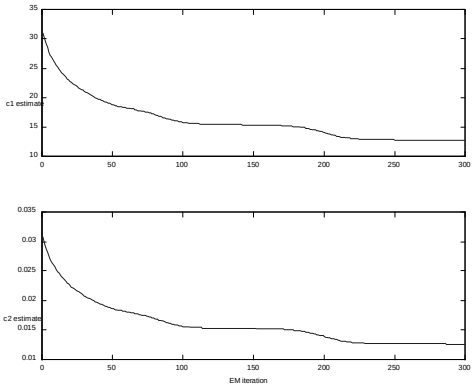

Figure 3. Estimates of rate parameters c1 and c2 as a function of EM iteration for a time-sampled Lotka reaction system.

EM gives reasonable estimates of the rate constants at the sampling rate selected, and an accurate indication of their relative magnitudes. Further investigation is required to determine the effect of varying the sampling rate and the accuracy of this approach for more complicated systems.

VI. CONCLUSION

Maximum likelihood techniques can be applied to time-series reactant concentration data to extract rate constant parameters. Reactants passing through more than one production or degradation channel present a hidden data problem that can be dealt with using the EM algorithm. Time sampled systems may be treated in the same manner. The estimation techniques developed can be applied to many biological systems of interest, and may allow for more accurate modeling of biological processes at the molecular level.

REFERENCES

[1] A. Arkin, J. Ross, and H.H. MacAdams, "Stochastic kinetic analysis of developmental -infected *Escherichia coli* cells," *Genetics* 1998, **149**:1633-1648.
[2] D.T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J.Comp.Phys.* 1976, **22**:403-434.
[3] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. New York: Wiley 2001.

ACKNOWLEDGEMENT

K.M. would like to thank NSERC for financial support of this work.