Stochastic Modeling in Systems Biology

Jinzhi Lei*

Zhou Pei-Yuan Center for Applied Mathematics
Tsinghua University, Beijing, 100084, China

October 4, 2010

Abstract

Many cellular behaviors are regulated by gene regulation networks, kinetics of which is one of the main subjects in the study of systems biology. Because of the low number molecules in these reacting systems, stochastic effects are significant. In recent years, stochasticity in modeling the kinetics of gene regulation networks have been drawing the attention of many researchers. This paper is a self contained review trying to provide an overview of stochastic modeling. I will introduce the derivation of the main equations for modeling the biochemical systems with intrinsic noise (chemical master equation, Fokker-Plan equation, reaction rate equation, chemical Langevin equation), and will discuss the relations between these formulations. The mathematical formulations for systems with fluctuations in kinetic parameters are also discussed. Finally, I will introduce the exact stochastic simulation algorithm and the approximate explicit tau-leaping method for making numerical simulations.

1 Introduction

Systems biology is an interdisciplinary science of discovering, modeling, understanding and ultimately engineering at the molecular level the dynamic relationships between the biological molecules that define living organisms1. It is increasingly hot in recent years as modeling molecular systems is not only fascinating but also possible in the post-genomics science. At the molecular level, many cellular behaviors are regulated by genetic regulation networks in which the stochasticity is significant. To describe these stochastic chemical kinetics, stochastic modeling is highlighted in recent years [19, 28, 50].

Chemical dynamics have been widely accepted to study chemical kinetics of reacting systems with large molecule populations, typically in the order of $10^{23}$. In these systems, the kinetics are nearly deterministic and can be described by a set of ordinary differential equations—the reaction rate equations, or partial differential equations if spatial movements are taken into account. Nevertheless, in intracellular molecule kinetics, stochasticity is significant because the numbers

*Email: jzei@mail.tsinghua.edu.cn

1According to Dr. Leroy Hood, president of the Institute of Systems Biology, Seattle. http://www.systemsbiology.org/Systems_Biology_in_Depth
of each molecule species are very low. For instance, only one gene, either active or inactive, is participate in most of gene expressions. In individual bacteria, no more than 20 transcriptions of mRNAs from a single gene. Typical numbers of same protein molecules in a cell are usually no more than a few thousands. Thus, fluctuations in protein activities are significant due to the low-number effect. Reaction rate equations fail to describe these fluctuations. In this review, I will introduce main equations for the stochastic modeling of such systems, and discuss numerical methods of stochastic simulations.

In this paper, I will first briefly review assumptions and notations for describing a biochemical system, followed by two examples of biological processes. Next, I will introduce several theoretical formulations of biochemical system kinetics with merely intrinsic noise, followed by the mathematical formulations for the situation in which kinetic parameters are random. Finally, some stochastic simulations methods supported by the previous theories are introduced. The paper will be concluded with a summarize of the theoretical structure of stochastic chemical kinetics.

2 Stochasticity in biological processes

It is no doubt that biological processes are essentially random. Cellular behavior and the cellular environment are stochastic. Phenotypes vary across isogenic populations and in individual cell over time. Gene expression is a fundamentally stochastic process, with randomness in transcription and translation leading to cell-to-cell variations in mRNA and protein levels. Noise propagation in gene networks has important consequences for cellular functions, being beneficial in some contexts and harmful in others.

Essentially, kinetics of biological molecules in living cells are consequences of chemically reacting systems. In this section, we first review basic assumption and description of general chemical systems, then give two examples to demonstrate their applications.

2.1 Chemical systems

Consider a system of well-stirred mixture of \( N \geq 1 \) molecular species \( \{S_1, \cdots, S_N\} \), inside some fixed volume \( \Omega \) and at constant temperature, through \( M \geq 1 \) reaction channels \( \{R_1, \cdots, R_M\} \). We can specify the dynamical state of this system by \( X(t) = (X_1(t), \cdots, X_N(t)) \), where

\[
X_i(t) = \text{the number of } S_i \text{ molecules in the system at time } t, \quad (t = 1, \cdots, N). \tag{1}
\]

We will described the evolution of \( X(t) \) from some given initial state \( X(t_0) = x_0 \). It is obvious that \( X(t) \) is a stochastic process, because the time at which a particular reaction occurs is random. Therefore, instead of tracking a single pathway, our goal it to study the evolution of the statistical properties of the system states.

In this review paper, we always assume that each reaction, once occur, completes instantaneously. This is to be distinguished with the systems involve reactions with delay. Further, we assume that the system is well stirred
such that at any moment, each reactions occur with equal probability at any position. Under these assumptions, each reaction channel $R_j$ associates with a propensity function $a_j$ and a state-change vector $v_j = (v_{ji}, \cdots, v_{jN})$, which are defined such that

$$a_j(x) dt = \text{the probability, given } X(t) = x, \text{ that one reaction } R_j \text{ will occur somewhere inside } \Omega \text{ in the next infinitesimal time interval } [t, t + dt), \ (j = 1, \cdots, M).$$

and

$$v_{ji} = \text{the change in the number of } S_i \text{ molecules produced by one } R_j \text{ reaction, } \ (j = 1, \cdots, M; i = 1, \cdots, N).$$

The propensity function and the state-change vector together specify the reaction channel $R_j$. Therefore, the equations below to describe the evolution of a biochemical system are derived from the propensity functions and stage-change vectors connected to the $M$ reaction channels.

The state-change vector is easy to be obtained from the reaction channel by counting the numbers of each molecule species that are consumed and produced in the reaction. For instance, if $R_1$ were the reaction $S_1 + 2S_2 \rightarrow 2S_3$, then $v_1 = (-1, -2, 2, \cdots)$.

Exact description of the propensity function associates with the \textit{ad hoc} stochasticity of deterministic chemical kinetics\cite{38}, and has a solid microphysical basis. In general, the function $a_j$ have the mathematical form\cite{17}

$$a_j(x) = c_j h_j(x).$$

Here $c_j$ is the specific probability rate constant for channel $R_j$, which is defined such that $c_j dt$ is the probability that a randomly chosen combination of $R_j$ reactant molecules will react accordingly in the next infinitesimal time interval $dt$. This probability $c_j dt$ equals the multiple of two parts, the probability of a randomly chosen combination of $R_j$ reactant molecules will collide in the next $dt$, and the probability that a colliding reactant molecules will actually react according to $R_j$. The first probability depends on the average relative speeds (which in turn depends on the temperature), the collision sections of a reactant molecules, and the system volume $\Omega$. The second probability depends on the chemical energy barrier $\Delta \mu$ of the reaction $R_j$ (Figure 1), and usually associate with temperature through a Boltzmann factor $e^{-\Delta \mu/k_B T}$, where $k_B$ is the Boltzmann constant.

The function $h_j(x)$ in (4) measures the number of distinct combinations of $R_j$ reactant molecules available in the state $x$. It can be easily obtained from the reaction $R_j$. For example, in the above reaction $R_1$, we would have $h_j(x) = x_1 x_2 (x_2 - 1)/2$. More examples are given below.

In general, for a chemical reaction

$$R_j : m_{j1}S_1 + \cdots + m_{jN}S_N \rightarrow n_{j1}S_1 + \cdots + n_{jN}S_N,$$

we would have

$$v_{ji} = n_{ji} - m_{ji}, \quad a_j(x) \propto \prod_{k=1}^N \frac{x_k!}{m_{jk}!(x_k - m_{jk})!}.$$
If for any $k$ and $j$, have $x_k \gg m_{jk}$, then approximately

$$a_j(x) = c_j \prod_{k=1}^{N} x_k^{m_{jk}}.$$

The reaction rate constant $c_j$ is deduced only from experiments, and usually depends on the system volume $\Omega$ and the temperature.

In real systems, most reaction channels are either monomolecular or bimolecular reaction. For a monomolecular reaction, the reaction rate constant $c_j$ is independent of the system volume $\Omega$. For a bimolecular reaction, the rate constant $c_j$ is inversely proportional $\Omega$. Trimolecular reactions do not physically occur in dilute solutions with appreciable frequency. One can consider a trimolecular reaction as the combined result of two bimolecular reactions, and involved an additional short-lived species. For such an “effective trimolecular” reaction, the approximate $c_j$ is proportional to $\Omega^{-2}$ [17].

### 2.2 Examples in gene regulations

#### 2.2.1 Gene expression

Gene expression is a basic biological process. Reactions in gene expression include promoter activity and inactivity, transcription, translation, and the decay of mRNA and proteins. Typical steps in gene expression are illustrated in Figure 2 (also refer [31]). Note that transcription is a process of mRNA synthesis as specified by the gene, in which only the information is read out and the gene is not consumed. Similarly, proteins are translated from the mRNA sequence according to the genetic code, and the mRNA is not consumed in this process.

In many gene regulations, the transition between active and inactive promoter states are regulated by a proteins (activator or repressor). In the case of activation, the activator bind to the inactive promoter to enhance the gene expression. The reaction channels $R_1$ and $R_2$ become

$$R_1 : X_5 + X_1 \rightarrow X_2, \quad R_2 : X_2 \rightarrow X_1 + X_5,$$

where $X_5$ stands for the number of the activator. The corresponding propensity functions and state-change vectors are $a_1(X) = c_1 X_1 X_5, a_2(X) = c_2 X_2, v_1 =$
Figure 2: (From ref. [31]) A model of gene expression. Each step represents the biochemical reactions which associate with transition between promoter states, production and decay of mRNAs and proteins (here $c_3 > c_4$).

$(-1,1,0,0,-1)$, and $v_2 = (1,-1,0,0,1)$. Similarly, in the case of repressor, we should have

$$R_1 : X_1 \rightarrow X_2 + X_5, \quad R_2 : X_5 + X_2 \rightarrow X_1,$$

and $a_1(X) = c_1 X_1$, $a_2(X) = c_2 X_2 X_5$, $v_1 = (-1,1,0,0,1)$, and $v_2 = (1,-1,0,0,-1)$.

### 2.2.2 Circadian oscillator

Figure 3 shows a simple model of circadian oscillator based on the common positive and negative control elements found experimentally [53]. In this model, two genes, an activator $A$ and a repressor $R$, which are transcribed into mRNA and subsequently translated into protein. The activator $A$ binds to the $A$ and $R$ promoters to increase their transcription rates. The protein $R$ binds to and sequester the protein $A$, and therefore acts as a repressor. Figure 3 shows the propensity functions, and stage-change vectors of this model (the $c_i$ in the Figure gives the rate constant of the reaction channel $R_i$). Note that in the reaction channel $R_{16}$, the complex breaks in to $R$ because of the degradation of $A$. Thus $R_{16}$ is not the reversion process of $R_{15}$.

### 3 Mathematical formulations–intrinsic noise

First, we assume that the propensity functions in previous are time independent, i.e., the reaction rates $c_j$ in (4) are constants. In this situation, the fluctuations in the system are inherent to the system of interest (intrinsic noise). The opposite case is extrinsic noise, which arises from variability in factors that are consider to be external. Mathematical formulations for extrinsic noise will be discussed in next section.

#### 3.1 Chemical master equation

From the propensity function given in previous, the state vector $X(t)$ is a jump-type Markov process on the non-negative $N$-dimensional integer lattice. In following analysis of such system, we will focus on the conditional probability function

$$P(x,t|x_0,t_0) = \text{Prob}\{X(t) = x, \text{ given that } X(t_0) = x_0\}. \quad (6)$$
Hereinafter, we use an upper case letter to denote a random variable, and the corresponding lower case letter for a possible value of that random variable.

Through the probability function, the average or expectation value of any quantity \( f(\mathbf{X}|\mathbf{x}_0, t_0) \) defined on the identical system is given by

\[
\langle f(\mathbf{X}|\mathbf{x}_0, t_0) \rangle = \sum \limits_{\mathbf{x}} f(\mathbf{x})P(\mathbf{x}, t|\mathbf{x}_0, t_0).
\]  

(7)

Physically, in an ensemble of identical systems starting from the same initial state \( \mathbf{X}(t_0) = \mathbf{x}_0 \), the function \( P(\mathbf{x}, t|\mathbf{x}_0, t_0) \) gives the fraction of subsystems with state \( \mathbf{X}(t) = \mathbf{x} \).

To derive a time evolution for the probability function \( P(\mathbf{x}, t|\mathbf{x}_0, t_0) \), we take a time increment \( dt \) and consider the variation between the probability of \( \mathbf{X}(t) = \mathbf{x} \) and of \( \mathbf{X}(t + dt) = \mathbf{x} \), given that \( \mathbf{X}(t_0) = \mathbf{x}_0 \). This variation is

\[
P(\mathbf{x}, t + dt|\mathbf{x}_0, t_0) - P(\mathbf{x}, t|\mathbf{x}_0, t_0) = \begin{cases} \text{Increasing of the probability in } dt, & \text{if } a_j(\mathbf{x} - \mathbf{v}_j), \\ \text{Decreasing of the probability in } dt, & \text{if } a_j(\mathbf{x} - \mathbf{v}_j), \end{cases}
\]

(8)

We take \( dt \) so small such that the probability of having two or more reactions in \( dt \) is negligible compared to the probability for only one reaction. Then increasing of the probability in \( dt \) occurs when a system with state \( \mathbf{X}(t) = \mathbf{x} \) reacts according to \( R_j \) in \( (t, t+dt) \), the probability of which is \( a_j(\mathbf{x} - \mathbf{v}_j)dt \).

Thus,

\[
\text{Increasing of the probability in } dt = \sum \limits_{j=1}^{M} P(\mathbf{x} - \mathbf{v}_j, t|\mathbf{x}_0, t_0)a_j(\mathbf{x} - \mathbf{v}_j)dt. \tag{9}
\]

Similarly, when a system with state \( \mathbf{X}(t) = \mathbf{x} \) reacts according to any reaction

Figure 3: (From ref. [53]) A biochemical network of the circadian oscillator model.
channel \( R_j \) in \((t, t + dt)\), the probability \( P(\mathbf{x}, t|\mathbf{x}_0, t_0) \) will decrease. Thus,

\[
\text{Decreasing of the probability in } dt = \sum_{j=1}^{M} P(\mathbf{x}, t|\mathbf{x}_0, t_0) a_j(\mathbf{x}) dt. \quad (10)
\]

Substituting (9) and (10) into (8), we obtain

\[
P(\mathbf{x}, t + dt|\mathbf{x}_0, t_0) - P(\mathbf{x}, t|\mathbf{x}_0, t_0) = \sum_{j=1}^{M} P(\mathbf{x} - \mathbf{v}_j, t|\mathbf{x}_0, t_0) a_j(\mathbf{x} - \mathbf{v}_j) dt
\]

\[
- \sum_{j=1}^{M} P(\mathbf{x}, t|\mathbf{x}_0, t_0) a_j(\mathbf{x}) dt.
\]

which yields, with the limit \( dt \to 0 \), the chemical master equation[17, 52]:

\[
\frac{\partial}{\partial t} P(\mathbf{x}, t|\mathbf{x}_0, t_0) = \sum_{|\mathbf{m}| \geq 1} \prod_{i=1}^{N} \left( \frac{-1}{m_i} \right) \frac{\partial^{m_i}}{\partial x_i^{m_i}} \left( A_{\mathbf{m}_1, \cdots, \mathbf{m}_N}(\mathbf{x}) P(\mathbf{x}, t) \right), \quad (11)
\]

The equation (11) is an exact consequence from the reaction channels characterized by propensity functions and stage-change vectors. If one can solve (11) for \( P \), we should be able to find out everything about the process \( X(t) \). However, such exact solution of (11) can rarely be obtained (refer [26, 49] for the examples of solving the chemical master equation analytically).

In fact, the chemical master equation (11) is a set of linear differential equations with constant coefficients. The difficult on solving the equation (11) comes from the high dimension, which equals the total number of possible states of the system under study. For example, for a system with 100 molecules species, each has two possible states \( (X_i(t) = 0 \) or \( 1) \), the system has totally \( 2^{100} \) possible states, and therefore the equation (11) contains \( 2^{100} \) equations! Even solving such a huge system numerically is a big challenge.

### 3.2 Fokker-Plank equation

If in a biochemical system, all components of \( X(t) \) are very large compared to 1, we can regard the components of \( X(t) \) as real numbers. We assume further that the functions \( f_j(\mathbf{x}) \equiv a_j(\mathbf{x}) P(\mathbf{x}, t|\mathbf{x}_0, t_0) \) are analytic in the real variable \( \mathbf{x} \). With these two assumptions, we can use Taylor’s expansion to write

\[
f_j(\mathbf{x} - \mathbf{v}_j) = f_j(\mathbf{x}) + \sum_{|\mathbf{m}| \geq 1} \prod_{i=1}^{N} \left( \frac{-1}{m_i} \right) \frac{\partial^{m_i}}{\partial x_i^{m_i}} \left( v_{ji}^{m_i} f_j(\mathbf{x}) \right). \quad (12)
\]

Here \( \mathbf{m} = (m_1, \cdots, m_N) \in \mathbb{Z}^N \), \( |\mathbf{m}| = m_1 + \cdots + m_N \). Substituting (12) into (11), we immediately obtain the chemical Kramer-Moyal equation (CKME) [17]

\[
\frac{\partial}{\partial t} P(\mathbf{x}, t) = \sum_{|\mathbf{m}| \geq 1} \prod_{i=1}^{N} \left( \frac{-1}{m_i} \right) \frac{\partial^{m_i}}{\partial x_i^{m_i}} \left( A^{m_1, \cdots, m_N}(\mathbf{x}) P(\mathbf{x}, t) \right), \quad (13)
\]

where

\[
A^{m_1, \cdots, m_N}(\mathbf{x}) = \sum_{j=1}^{M} v_{j1}^{m_1} \cdots v_{jN}^{m_N} a_j(\mathbf{x}).
\]
Hereinafter, we omit the initial condition \((x_0, t_0)\). The chemical Kramer-Moyal equation is a “semi-rigorous” consequence of the chemical master equation (11).

If we truncate the right hand side at \(|m| = 2\), we obtain the chemical Fokker-Plank equation (CFPE)

\[
\frac{\partial}{\partial t} P(x, t) = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} A_i(x) P(x, t) + \frac{1}{2} \sum_{1 \leq i, k \leq N} \frac{\partial^2}{\partial x_i \partial x_k} B_{ik}(x) P(x, t). \tag{14}
\]

where

\[
A_i(x) = \sum_{j=1}^{M} v_{ji} a_j(x), \quad B_{ik}(x) = \sum_{j=1}^{M} v_{ji} v_{jk} a_j(x). \tag{15}
\]

We will re-obtain the chemical Fokker-Plank equation below from the chemical Langevin equation.

### 3.3 Reaction rate equation

If we multiply the chemical master equation (11) through by \(x_i\), sum over all \(X\), we obtain the following chemical ensemble average equation (CEAE)

\[
\frac{d\langle X_i \rangle}{dt} = \sum_{j=1}^{M} v_{ji} \langle a_j(X) \rangle \quad (i = 1, 2, \cdots, N). \tag{16}
\]

Equation (16) is an exact consequence of (11). Nevertheless, (16) is not a complete system unless all reaction channels are monomolecular. In this situation, the propensity functions \(a_j(X)\) are linear, and therefore

\[
\langle a_j(X) \rangle = a_j(\langle X \rangle) \quad (j = 1, \cdots, N). \tag{17}
\]

Thus, the populations evolve deterministically according to a set of ordinary differential equations

\[
\frac{dx_i}{dt} = \sum_{j=1}^{M} v_{ji} a_j(x) \quad (i = 1, \cdots, N), \tag{18}
\]

where the components of \(x(t)\) are now real variables.

In usual, the mean populations do not evolve according to (18) when there are higher order reactions. However, (18) is sometimes heuristic if we assuming that the fluctuations are not important and (17) holds approximately. Consequently, the ordinary differential equations (18) also hold under this assumption. The equation (18) is referred as the macroscopic reaction rate equation (RRE) or chemical rate equation.

The reaction rate equation (18) is often written in terms of the species concentrations

\[
Z_i(t) \equiv \frac{X_i(t)}{\Omega} \quad (i = 1, \cdots, N).
\]

The “concentrations" form of the reaction rate equation has form

\[
\frac{dz_i}{dt} = \sum_{j=1}^{M} v_{ji} \tilde{a}_j(z) \quad (i = 1, \cdots, N), \tag{19}
\]
where
\[ \tilde{a}_j(z) = a_j(\Omega z)/\Omega. \]

If we examine the \( \Omega \) dependence of the propensity functions for monomolecular, bimolecular, and trimolecular reactions, we find that the function \( \tilde{a}_j \) is functionally identical to \( a_j \) except that the rate constants \( c_j \) has been replaced by the reaction rate constant \( k_j \).

The reaction rate equations are most commonly used in modeling biochemical reacting systems. However, this equation fails to describe the stochastic effects which can be very important in biological processes. We will introduce the chemical Langevin equation below that has a facility to describe the stochasticity, and easy to be studied, at least numerically.

### 3.4 Chemical Langevin equation

The chemical Langevin equation was derived leads to an approximatetime-evolution equation of the Langevin type. The derivation of the equation, given by Gillespie, is based on the chemical master equation and two explicit dynamical conditions as detailed below. Most of the following refer to Gillespie’s original paper [17].

Suppose the system’s state at the current time \( t \) is \( X(t) = x \). Let \( K_j(x, \tau) \) \( (\tau > 0) \) be the number of \( R_j \) reactions that occur in the subsequent time interval \([t, t+\tau]\). Since each of those reactions will change the \( S_i \) population by \( v_{ji} \), the number of \( S_i \) molecules in the system at time \( t+\tau \) will be

\[
X_i(t+\tau) = x_i + \sum_{j=1}^{M} K_j(x, \tau)v_{ji}, \quad (i = 1, \cdots, N). \tag{20}
\]

We note that \( K_j(x, \tau) \) is a random variable, and therefore \( X_i(t+\tau) \) is random.

We will obtain an approximation of \( K_j(x, \tau) \) below by imposing the following conditions:

**Condition (i)** Require \( \tau \) to be small enough such that the change in the state during \([t, t+\tau]\) will be so slight that none of the propensity functions changes its value “appreciably”.

**Condition (ii)** Require \( \tau \) to be large enough that the expected number of occurrences of each reaction channel \( R_j \) in \([t, t+\tau]\) be much larger than 1.

From condition (i), the propensity functions satisfy
\[
a_j(x'(t)) \approx a_j(x), \quad \forall t' \in [t, t+\tau], \forall j \in [1, M]. \tag{21}
\]

Thus, the probability of the reaction \( R_j \) to occur in any infinitesimal interval \( d\tau \) within \([t, t+\tau]\) is \( a_j(x)d\tau \). Thus, \( K_j(x, \tau) \), the occurrence of “events” of reaction \( R_j \) in the time interval \([t, t+\tau]\), will be a statistically independent Poisson random variable, and is denoted by \( P_j(a_j(x), \tau) \). So (20) can be approximated by
\[
X_i(t+\tau) = x_i + \sum_{j=1}^{M} v_{ji}P_j(a_j(x), \tau), \quad (i = 1, \cdots, N) \tag{22}
\]
according to condition (i).

The mean and variance of $P(a, \tau)$ are

$$\langle P(a, \tau) \rangle = \text{var}(P(a, \tau)) = a\tau.$$  

(23)

Thus, condition (ii) means

$$a_j(x)\tau \gg 1, \quad \forall j \in [1, M].$$  

(24)

The inequality (24) allows us to approximate each Poisson random variable by a *normal* random variable with the same mean and variance. This leads to further approximation

$$X_i(t + \tau) = x_i + \sum_{j=1}^{M} v_{ji} N_j(a_j(x)\tau, a_j(x)\tau), \quad (i = 1, \cdots, N)$$  

(25)

where $N_j(m, \sigma^2)$ denotes the normal random variable with mean $m$ and variance $\sigma^2$. Here the $M$ random variables are independent to each other. Notice that in the above approximation, we have converted the molecular population $X_i$ from discretely changing integers to continuously changing real variables.

We note the linear combination theorem for normal random variables,

$$N(m, \sigma^2) = m + \sigma N(0, 1),$$  

(26)

the equation (25) can be rewritten as

$$X_i(t + \tau) = x_i + \sum_{j=1}^{M} v_{ji} a_j(x)\tau + \sum_{j=1}^{M} v_{ji} a_j(x)\tau^{1/2} N_j(0,1) \quad (i = 1, \cdots, N).$$  

(27)

Now, we are ready to obtain Langevin type equations by making some purely notational changes. First, we denote the time interval $\tau$ by $dt$, and write

$$dX_i = X_i(t + dt) - X_i(t).$$

Next, introduce $M$ temporally uncorrelated, independent random process $W_j(t)$, satisfying

$$dW_j(t) = W_j(t + dt) - W_j(t) = N_j(0,1)(dt)^{1/2} \quad (j = 1, \cdots, N).$$  

(28)

It is easy to verify that the processes $W_j(t)$ have *stationary independent increments with mean 0*, i.e.,

$$\langle dW_j(t) \rangle = 0, \quad \langle dW_i(t)dW_j(t') \rangle = \delta_{ij}\delta(t - t')dt, \quad \forall 1 \leq i, j \leq M, \forall t, t'.$$  

(29)

Thus, each $W_j$ is a *Wiener process* (or referred to as *Brownian motion*). Finally, recalling that $x$ stands for $X(t)$, the equation (25) becomes a *Langevin equation* (or *stochastic differential equation*)

$$dX_i = \sum_{j=1}^{M} v_{ji} a_j(X)dt + \sum_{j=1}^{M} v_{ji} a_j^{1/2}(X)dW_j \quad (i = 1, \cdots, N).$$  

(30)
The stochastic differential equation (30) is the desired chemical Langevin equation (CLE). The solution of (30) with initial condition $X(0) = X_0$ is a stochastic process $X(t)$ satisfying

$$X(t) = X_0 + \sum_{j=1}^{M} \int_0^t v_{ji} a_j(X(s)) ds + \sum_{j=1}^{M} \int_0^t v_{ji} a_j^{1/2}(X(s)) dW_j(s).$$

(31)

Here Itô integral is used as the stochastic fluctuations are intrinsic[52].

In the above, it is obvious that the conditions (i) and (ii) are contradict to each other. It may be very well happen that both conditions cannot be satisfied simultaneously. In this case, the derivation of the chemical Langevin equation may fail. But there are many practical circumstances in which the two conditions can be simultaneously satisfied. As the inequality (24) implies that $a_j(x)$ is large when $\tau$ is small enough as required by condition (i). This is possible when the system has large molecular populations for each molecule species. Since $a_j(x)$ is typically proportional to one or more components of $x$.

Even when the conditions (i) and (ii) cannot be satisfied simultaneously, the chemical Langevin equation (30) is still useful. As we will discuss below.

### 3.5 Validity of the chemical Langevin equation

In many intracellular biochemical systems such as gene expressions and genetic networks, the molecule populations are small, and the reactions are usually slow. In these systems, the derivation of the chemical Langevin equation may fail. Nevertheless, the equation (30) is ‘statistically correct’ for such systems as it provides correct time-evolutions of the statistical behavior. The reasons are given below.

#### 3.5.1 Time-evolution of the mean

If we take average to both sides of (30), noticing

$$\langle a_j(X)^{1/2}dW_j \rangle = 0 \quad (j = 1, \cdots, M)$$

from the Itô interpretation, we have

$$\frac{d\langle X_i \rangle}{dt} = \sum_{j=1}^{M} v_{ji} \langle a_j(X) \rangle \quad (i = 1, \cdots, N).$$

(32)

This gives the same chemical ensemble average equation (16) as we have obtained from the chemical master equation.

The above arguments show that the chemical Langevin equation and the chemical master equation give the same time-evolution of the mean populations.

#### 3.5.2 Time-evolution of the correlations

The correlations between the molecule numbers are defined as

$$\sigma_{ij}(t) = \langle X_i(t)X_j(t) \rangle - \langle X_i(t) \rangle \langle X_j(t) \rangle, \quad (1 \leq i, j \leq N).$$

(33)
If we multiply the chemical master equation (11) through by \(x_i x_j\), sum over all \(x\), we can obtain

\[
\frac{d\langle X_i X_j \rangle}{dt} = \sum_{k=1}^{M} \left[ \langle X_j v_{ki} a_k(X) \rangle + \langle X_i v_{kj} a_k(X) \rangle + \langle v_{ki} v_{kj} a_k(X) \rangle \right].
\] (34)

Furthermore, (16) gives

\[
\frac{d\langle X_i \rangle \langle X_j \rangle}{dt} = \sum_{k=1}^{M} \left[ \langle v_{ki} a_k(X) \rangle \langle X_j \rangle + \langle v_{kj} a_k(X) \rangle \langle X_i \rangle \right].
\] (35)

Thus, recalling (15), we have the chemical ensemble correlations equation (CECE)

\[
\frac{d\sigma_{ij}}{dt} = \langle A_i(X)(X_j - \langle X_j \rangle) \rangle + \langle (X_i - \langle X_i \rangle) A_j(X) \rangle + \langle B_{ij}(X) \rangle.
\] (36)

The equation (36) is exact from the chemical master equation.

Now, starting from the chemical Langevin equation (30) and applying the Itô formula [37], we have

\[
d(X_i X_j) = X_i(dX_j) + X_j(dX_i) + (dX_i)(dX_j) = \sum_{k=1}^{M} \left[ X_j v_{ki} a_k(X) + X_i v_{kj} a_k(X) + v_{ki} v_{kj} a_k(X) \right] dt + \sum_{k=1}^{M} \left[ X_i v_{ki} a_k^{1/2}(X) + X_j v_{kj} a_k^{1/2}(X) \right] dW_k + o(dt)
\]

Taking the average, we re-obtain (34), which yields the same chemical ensemble correlations equation (36).

Therefore, the chemical Langevin equation and the chemical master equation give the same equation for the time-evolution of the correlations.

At near equilibrium state, if the random fluctuations are not important, we have approximately

\[
A_i(X) \approx A_i(\langle X \rangle) + \sum_{l=1}^{N} \frac{\partial A_i(\langle X \rangle)}{\partial X_l} (X_l - \langle X_l \rangle), \quad \langle B_{ij}(X) \rangle \approx B_{ij}(\langle X \rangle).
\]

Substituting above approximations into (36), and defining the matrices

\[
\sigma = (\sigma_{ij}), \quad A = \left( \frac{\partial A_i(\langle X \rangle)}{\partial X_l} \right), \quad B = \langle B_{ij}(\langle X \rangle) \rangle,
\] (37)

we have

\[
\frac{d\sigma}{dt} = (A\sigma + \sigma A^T) + B.
\] (38)

Equation (38) gives the linear approximation of the time-evolution of the correlation functions near equilibrium. This is an example of the famous fluctuation-dissipation theorem [52].
3.5.3 Fokker-Plank equation

Consider the chemical Langevin equation \( (30) \), the solution \( X(t) \) satisfies

\[
X_i(t + \Delta t) = X_i(t) + \sum_{k=1}^{M} v_{ki}a_j(X)\Delta t + \sum_{k=1}^{M} v_{ki}a_k^{1/2}(X)\Delta W_k(t), \quad (i = 1, \ldots, N)
\]

where

\[
\Delta W_k(t) = W_k(t + \Delta t) - W_k(t).
\]

Thus, assuming \( X(t) = x \), the displacement \( \Delta x = X(t + \Delta t) - X(t) \) satisfies

\[
\langle \Delta x_i \rangle = \sum_{k=1}^{M} v_{ki}a_k(x)\Delta t
\]

and

\[
\langle \Delta x_i \Delta x_j \rangle = \sum_{k=1}^{M} v_{ki}v_{kj}a_k(x)\Delta t.
\]

Let \( P(x, t) \) to be the conditioned probability density function as previous, then \( P(x, t) \) satisfies

\[
P(x, t + dt) - P(x, t) = \int_{\Delta x \in \mathbb{R}^N} P(x - \Delta x, t)W(\Delta x, dt; x - \Delta x, t)d\Delta x
\]

\[
- \int_{\Delta x \in \mathbb{R}^N} P(x, t)W(\Delta x, dt; x, t)d\Delta x,
\]

where \( W(\Delta x, dt; x, t) \) is the transition probability from \( X(t) = x \) to \( X(t + dt) = x + \Delta x \). Applying Taylor expansion to the function \( h(x - \Delta x) \equiv P(x - \Delta x, t)W(\Delta x, dt; x - \Delta x, t) \), noticing

\[
\int_{\Delta x \in \mathbb{R}^N} \Delta x_i W(\Delta x, dt; x, t) = \langle \Delta x_i \rangle
\]

and

\[
\int_{\Delta x \in \mathbb{R}^N} \Delta x_i \Delta j W(\Delta x, dt; x, t)d\Delta x = \langle \Delta x_i \Delta x_j \rangle,
\]

and letting \( dt \to 0 \), we re-obtain the chemical Fokker-Plank equation \( (14) \) as previous.

Thus, the chemical Langevin equation and the chemical master equation yield the same “semi-rigorous” chemical Fokker-Plank equation.

4 Mathematical formulations–fluctuation in kinetic parameters

In the previous discussion, we assume the reaction rates \( c_j \) in \( (4) \) are constants. This is only a rough approximation of the real world, in which cell environments are stochastic. In this situation, the reaction rates are random. Here, we will discuss the mathematical formulations for situations in which there are fluctuations in kinetic parameters.
4.1 Reaction rate as a random process

When the reaction rate \( c_j \) is random, the previous propensity function \( a_j(x) \) in (4) should be rewritten as

\[
a_j(x, t) = c_j(t)h_j(x).
\]  

(40)

Here \( c_j(t) \), as previous, is the specific probability reaction rate for channel \( R_j \) at time \( t \). This rate is usually random, depending on the fluctuating environment whose explicit time dependence is not known. A reasonable approximation is given below.

In previous discussions, we have \( c_j \propto e^{-\Delta \mu/k_B T} \). If there are noise perturbations to the energy barrier, we can replace \( \Delta \mu \) by \( \Delta \bar{\mu} - k_B T \eta(t) \), where \( \eta(t) \) is a stochastic process. Consequently, we can replace the reaction rate \( c_j(t) \) by

\[
c_j(t) = \bar{c}_j e^{\eta(t)} / \langle e^{\eta(t)} \rangle,
\]

(41)

where \( \bar{c}_j \) is a constant, measures the mean of the reaction rate.

In many cases, the noise perturbation \( \eta(t) \) can be described by an Ornstein-Uhlenbeck process, which is given by a solution of following stochastic differential equation

\[
d\eta = -\eta/\tau dt + \sigma \sqrt{2/\tau} dW.
\]

(42)

Here \( W \) is a Wiener process, \( \tau \) and \( \sigma \) are positive constants, measuring the autocorrelation time and variance, respectively. It is easy to verify that \( \eta(t) \) is normally distributed and has an exponentially decaying stationary autocorrelation function[52, 16]

\[
\langle \eta(t)\eta(t') \rangle = \sigma^2 e^{-|t-t'|/\tau}.
\]

(43)

The Ornstein-Uhlenbeck process is an example of color noise.

With \( \eta(t) \) an Ornstein-Uhlenbeck process, the stationary distribution of the reaction rate \( c_j(t) \) is then log-normal. Log-normal distribution have been seen in many applications[33]. For instance, log-normal rather than normal distribution have been measured for gene expression rates[44].

When \( \sigma \) is small, we have approximately \( \langle e^{\eta(t)} \rangle = e^{\sigma^2/2} \), therefore the reaction rate (41) becomes

\[
c_j(t) = \bar{c}_j e^{\eta(t) - \sigma^2/2}.
\]

(44)

Thus, the propensity function (40) now becomes

\[
a_j(x, t) = \bar{c}_j e^{\eta(t) - \sigma^2/2} h_j(x),
\]

(45)

with \( \eta(t) \) satisfies (42).

Next, we will introduce generalizations of the reaction rate equation and the chemical Langevin equation to describe the reaction system with fluctuation in reaction rates. We note that in this situation, the state variable \( X \) alone is not enough to describe the system state, and the environment state has to be taken into account as well. Thus, to generalize the chemical master equation or the Fokker-Plank equation, we should replace the previous probability function by \( P(x, c, t|x_0, c_0, t) \), where \( c = \{c_1, \cdots, c_M\} \). This is omitted here. In the following formulations, we always assume that \( \sigma \) is small, and (45) is applied.
4.2 Reaction rate equation

Substituting the propensity functions of form (45) and (42) into the reaction rate equation (18), we obtain the following stochastic differential equations

\[ dX_i = \left( \sum_{j=1}^{M} v_{ji} \hat{c}_j e^{\eta_j - \sigma_j^2/2} h_j(X) \right) dt, \quad (i = 1, \ldots, N) \]  
\[ d\eta_j = -\left( \eta_j / \tau_j \right) dt + \sigma_j \sqrt{2/\tau_j} dW_j, \quad (j = 1, \ldots, M) \]

Here \( \hat{c}_j, \tau_j, \sigma_j \) are positive constants. The equations (46)-(47) generalize the reaction rate equation to describe the reaction system with fluctuations in the kinetic parameters.

4.3 Chemical Langevin equation

Similar to the above strategy, substituting the propensity functions of form (45) and (42) into the chemical Langevin equation (30), we obtain the following stochastic differential equations

\[ dX_i = \left( \sum_{j=1}^{M} v_{ji} \hat{c}_j e^{\eta_j - \sigma_j^2/2} h_j(X) \right) dt + \sum_{j=1}^{M} v_{ji} \left( \hat{c}_j e^{-\eta_j - \sigma_j^2/2} h_j(X) \right)^{1/2} dW_j, \quad (i = 1, \ldots, N) \]
\[ d\eta_j = -\left( \eta_j / \tau_j \right) dt + \sigma_j \sqrt{2/\tau_j} dW'_j, \quad (j = 1, \ldots, M) \]

Here \( W'_j \) are independent Wiener processes, and \( \hat{c}_j, \tau_j, \sigma_j \), as previous, are positive constants. This set of generalized chemical Langevin equations describes the stochasticity of chemical systems with both intrinsic noise and fluctuations in kinetic parameters.

5 Stochastic simulations

We have introduced several equations for modeling systems of biochemical reactions. Here, we will introduce stochastic simulation methods that intend to mimic a random process \( X(t) \) of system evolution. Once we have enough sampling pathways of the random process, we are able to calculate the probability density function \( P(x, t) \) and other statistical behaviors including the mean trajectory and correlations.

5.1 Stochastic simulation algorithm

Assume the system is in state \( x \) at time \( t \). We define the next-reaction density function \( p(\tau, j|x, t) \) as

\[ p(\tau, j|x, t) = \frac{\text{probability that, given } X(t) = x, \text{ the next reaction in } \Omega \text{ will occur in the infinitesimal time interval } [t + \tau, t + \tau + d\tau], \text{ and will be an } R_j \text{ reaction.}}}{R_j} \]  

(50)
Then an elementary probability argument based on the propensity function (2) yields [14, 15]

\[ p(\tau, j|x, t) = a_j(x) \exp(-a_0(x)\tau), \quad (0 \leq \tau < \infty, j = 1, \ldots, M) \] (51)

where

\[ a_0(x) = \sum_{k=1}^{M} a_k(x). \]

This formula provides the basis for the stochastic simulation algorithm (SSA) (also known as the Gillespie algorithm).

Following direct method is perhaps the simplest to generate a pair of numbers \((\tau, j)\) in accordance with the probability (51) [15]: First generate two random numbers \(r_1\) and \(r_2\) from the uniform distribution in the unit interval, and then take

\[ \tau = (1/a_0(x)) \ln(1/r_1) \] (52)

\[ j = \text{the smallest integer satisfying } \sum_{j'=1}^{j} a_{j'}(x) > r_2 a_0(x). \] (53)

Below are main steps in the stochastic simulation algorithm (for details, refer [15])

1. **Initialization**: Let \(X = x_0\), and \(t = t_0\).

2. **Monte Carlo step**: Generate a pair of random numbers \((\tau, j)\) according to the probability density function (51) (replace \(x\) by \(X\)).

3. **Update**: Increase the time by \(\tau\), and replace the molecule count by \(X + v_j\).

4. **Iterate**: Go back to Step 2 unless the simulation time has been exceeded.

The stochastic simulation algorithm consists with the chemical master equation since (11) and (51) are exact consequence of the propensity function (2). This algorithm numerically simulation the time evolution of a given chemical system, and gives a sample trajectory of the real system.

**5.2 Tau-leaping algorithm**

In the derivation of the chemical Langevin equation, we can always select \(\tau\) small enough such that the the following leap condition is satisfied: During \([t, t + \tau)\), no propensity function is likely to change its value by a significant amount. Consequently, the previous arguments indicate that we can approximately leap the system by a time \(\tau\) by taking

\[ X(t + \tau) \approx x + \sum_{j=1}^{M} P_j(a_j(x), \tau)v_j, \] (54)

where \(x = X(t)\).

The equation (54) is the basic of the tau-leaping algorithm [18]: Starting from the current state \(x\), we first choose a value \(\tau\) that satisfies the leap condition.
Next, we generate for each \( j \) a random number \( k_j \) according to Poisson distribution with mean \( a_j(x)\tau \). Finally, we update the state from \( x \) to \( x + \sum_j k_j v_j \), and increase the time by \( \tau \).

There are two practical issues need to be resolved to effectively using the tau-leaping algorithm: First, how can we estimate the largest value of \( \tau \) that satisfies the leap condition? Second, how can we ensure that the generated \( k_j \) values do not result in negative populations?

The original estimation of the largest value of \( \tau \) is given by Gillespie as [18]

\[
\tau = \min_{j \in [1,M]} \left\{ \varepsilon a_0(x) / \left| \sum_{i=1}^{N} \xi_i(x) b_{ji}(x) \right| \right\},
\]

(55)

where \( 0 < \varepsilon \ll 1 \) is a accuracy-control parameter, and

\[
a_0(x) = \sum_{j=1}^{M} a_j(x), \quad \xi_i(x) = \sum_{j=1}^{M} a_j(x) v_{ji}, \quad b_{ji}(x) = \frac{\partial a_j(x)}{\partial x_i}.
\]

(56)

In [20] and [6], two successive refinements were made. The latest \( \tau \)-selection procedure [6] is more accurate, easier to code, and faster to execute than the earlier procedures, but logically more complicated.

If the \( \tau \) value generated above is much larger than the time required for the stochastic simulation algorithm, this approximate procedure will be faster than the exact stochastic simulation algorithm. However, it \( \tau \) turns out to be less than a few multiples of the time required for the stochastic simulation algorithm to make an exact time step (in the order of \( 1/a_0(x) \)), then it would be better to use the stochastic simulation algorithm.

To avoid the negative populations in tau-leaping, several strategies have been proposed, in which the unbounded Poisson random numbers \( k_j \) are replaced by bonded binormal random numbers [7, 51]. In 2005, Cao et al. [3] proposed a new approach to resolve this difficulty. In this new Poisson tau-leaping procedure, the reaction channels are separated into two classes: critical reactions that may exhaust one of its reactants after some firings, and noncritical reactions otherwise. Next, the noncritical reactions are handled by the regular tau-leaping method to obtain a leap time \( \tau' \). And apply the exact stochastic simulation algorithm to the critical reactions. This gives the time \( \tau'' \), and the index \( j_c \) of the next critical reactions. The actual time step \( \tau \) is then taken to be the smaller of \( \tau' \) and \( \tau'' \). If the former, no critical reactions fires, and if the latter, only one critical reactions \( R_{j_c} \) fires.

Below are main steps for the modified tau-leaping procedure that avoids negative populations (refer [3] for details)

1. **Initialization**: Let \( X = x_0 \), and \( t = t_0 \).

2. **Identify the critical reactions**: Identify the reaction channels \( R_j \) for which \( a_j(X) > 0 \) and may exhaust one of its reactants after some firings.

3. **Calculate the leap time**: Compute the largest leap time \( \tau' \) for the noncritical reactions.

4. **Monte Carlo step**: Generate \( (\tau'', j_c) \) for the next critical reaction according to the modified density function according to (51).
5. Determine next step:

(a) If $\tau' < \tau''$: Take $\tau = \tau'$. For all critical reactions $R_i$, set $k_i = 0$. For all the noncritical reactions $R_j$, generate $k_j$ as a Poisson random variable with mean $a_j(X)\tau$.

(b) If $\tau'' \leq \tau'$: Take $\tau = \tau''$. Set $k_{j_c} = 1$, and for all other critical reactions set $k_j = 0$. For all the noncritical reactions $R_j$, generate $k_j$ as a Poisson random variable with mean $a_j(X)\tau$.

6. **Update**: Increase the time by $\tau$, and replace the molecular count by $X + \sum_{j=1}^{M} k_j v_j$.

7. **Iterate**: Go back to Step 2 unless the simulation time has been exceeded.

5.3 Other simulation methods

In addition to the prominent approximate acceleration procedure tau-leaping, there are some other strategies that tend to speedup the stochastic simulation algorithm. Here we briefly outline two of the most promising methods.

Many real systems in biological processes involve chemical reactions with different time scales, “fast” reactions fire very much more frequently than “slow” ones. Procedures to handle such systems often involve a stochastic generalization of the quasi-steady-state assumption or partial (rapid) equilibrium methods of deterministic chemical kinetics [2, 5, 22, 42, 46]. The slow-scale stochastic simulation algorithm (ssSSA) (or multiscale stochastic simulation algorithm) is a systematic procedure for partitioning the system into fast and slow reactions, and only simulate the slow reactions by specially modified propensity functions [2, 4, 5].

Another approach to simulate multiscale chemical reaction systems include different kinds of Hybrid methods [1, 12, 23, 45]. Hybrid methods combine the deterministic reaction rate equation with the stochastic simulation algorithm. The idea is to split the system into two regimes: the continuous regime of large molecule population species, and the discrete regime of small molecule population species. The continuous regimes is treated by ordinary differential equations, while the discrete regime is simulated by the stochastic simulation algorithm. Hybrid methods efficiently utilize the multiscale properties of the system. However, because of lacking a rigorous theoretical foundation, there are still many unsolved problems [32].

There are many approaches trying to find the numerical solution of the chemical master equation [8, 10, 11, 24, 25, 27, 29, 36]. In addition, numerical methods for the Langevin equation have been well documented (refer [30] for example). We will not get into these two subjects here.

6 Summary

In modeling a well-stirred chemically reacting systems, the chemical master equation provides ‘exact’ description of the time evolution of the states. However, it is difficulty to directly study the chemical master equation because of the dimension problem. Several approximations are therefore developed, including the Fokker-Plank equation, reaction rate equation, and chemical Langevin
The reaction rate equation is widely used when fluctuations are not important. When the fluctuations are significant, the chemical Langevin equation can provide reasonable description for the statistical properties of the kinetics, despite the conditions in deriving the chemical Langevin equation may not hold.

When there are noise perturbations to the kinetic parameters, there is no simple way to model the system dynamics because the time dependence of environment variables can be very complicated. In a particular case, we can replace the reaction rates by log-normal random variables, and generalize the reaction rate equation or the chemical Langevin equation to describe the dynamics of a chemical system with extrinsic noise.

Stochastic simulation algorithm (SSA) is an ‘exact’ numerical simulation that shares the same fundamental basis as the chemical master equation. The approximate explicit tau-leaping produce, on the other hand, is closely related to the chemical Langevin equation. The robustness and efficiency of the two methods have been considerable improved in recent years, and these procedures seems to be nearing maturity [19]. In the last few years, some other strategies have been developed for simulating the systems that are dynamically stiff [19, 32].

We conclude this paper with Figure 4, which summarizes the theoretical structure of stochastic chemical kinetics (also refer [19]).

References


Figure 4: (Modified from ref. [19]) Theoretical structure of stochastic chemical kinetics. Everything originate from the fundamental premise of propensity function given at the top box. Solid arrows show the exact inferences routes, and dotted arrows show the approximate ones, with justified conditions in braces. Solid boxes are exact results: the chemical master equation (CME), the stochastic simulation algorithm (SSA), the chemical ensemble coefficient equation (CECE), and the chemical ensemble average equation (CEAE). Dashed boxes are approximate results: the Tau-leaping algorithm, the chemical Langevin equation (CLE), the chemical Kramers-Moyal equation (CKME), the chemical Fokker-Plank equation (CFPE), and the reaction rate equation (RRE).


