

---

---

# Biomolecular Feedback Systems

---

Domitilla Del Vecchio  
MIT

Richard M. Murray  
Caltech

DRAFT v0.4a, January 16, 2011  
© California Institute of Technology  
All rights reserved.

This manuscript is for review purposes only and may not be reproduced, in whole or in part, without written consent from the authors.

---

---

## Chapter 4

### Stochastic Modeling and Analysis

In this chapter we explore stochastic behavior in biomolecular systems, building on our preliminary discussion of stochastic modeling in Section 2.1. We begin by reviewing the various methods for modeling stochastic processes, including the chemical master equation (CME), the chemical Langevin equation (CLE) and the Fokker-Planck equation (FPE). Given a stochastic description, we can then analyze the behavior of the system using a variety of stochastic simulation and analysis tools. In many cases, we must simplify the dynamics of the system in order to obtain a tractable model, and we describe several methods for doing so, including finite state projection, linearization and Markov chain representations. We also investigate how to use data to identify some the structure and parameters of stochastic models.

*Prerequisites.* This chapter makes use of a variety of topics in stochastic processes that are not covered in AM08. Readers should have a good working knowledge of basic probability and some exposure to simple stochastic processes (e.g., Brownian motion), at the level of the material presented in Appendix C (drawn from [53]).

#### 4.1 Stochastic Modeling of Biochemical Systems

Chemical reactions in the cell can be modeled as a collection of stochastic events corresponding to chemical reactions between species, including binding and unbinding of molecules (such as RNA polymerase and DNA), conversion of one set of species into another, and enzymatically controlled covalent modifications such as phosphorylation. In this section we will briefly survey some of the different representations that can be used for stochastic models of biochemical systems, following the material in the textbooks by Phillips *et al.* [56], Gillespie [28] and Van Kampen [43].

##### Statistical physics

At the core of many of the reactions and multi-molecular interactions that take place inside of cells is the chemical physics associated with binding between two molecules. One way to capture some of the properties of these interactions is through the use of statistical mechanics and thermodynamics.

As described briefly already in Chapter 2, the underlying representation for both statistical mechanics and chemical kinetics is to identify the appropriate microstates of the system. A microstate corresponds to a given configuration of the components (species) in the system relative to each other and we must enumerate all possible configurations between the molecules that are being modeled.

In statistical mechanics, we model the configuration of the cell by the probability that system is in a given microstate. This probability can be calculated based on the energy levels of the different microstates. Consider a setting in which our system is contained within a reservoir. The total (conserved) energy is given by  $E_{\text{tot}}$  and we let  $E_r$  represent the energy in the reservoir. Let  $E_s^{(1)}$  and  $E_s^{(2)}$  represent two different energy levels for the system of interest and let  $W_r(E_r)$  be the number of possible microstates of the reservoir with energy  $E_r$ . The laws of statistical mechanics state that the ratio of probabilities of being at the energy levels  $E_s^{(1)}$  and  $E_s^{(2)}$  is given by the ratio of number of possible states of the reservoir:

$$\frac{P(E_s^{(1)})}{P(E_s^{(2)})} = \frac{W_r(E_{\text{tot}} - E_s^{(1)})}{W_r(E_{\text{tot}} - E_s^{(2)})}. \quad (4.1)$$

Defining the entropy of the system as  $S = k_B \ln W$ , we can rewrite equation (4.1) as

$$\frac{W_r(E_{\text{tot}} - E_s^{(1)})}{W_r(E_{\text{tot}} - E_s^{(2)})} = \frac{e^{S_r(E_{\text{tot}} - E_s^{(1)})/k_B}}{e^{S_r(E_{\text{tot}} - E_s^{(2)})/k_B}}.$$

We now approximate  $S_r(E_{\text{tot}} - E_s)$  in a Taylor series expansion around  $E_{\text{tot}}$ , under the assumption that  $E_r \gg E_s$ :

$$S_r(E_{\text{tot}} - E_s) \approx S_r(E_{\text{tot}}) - \frac{\partial S_r}{\partial E} E_s.$$

From the properties of thermodynamics, if we hold the volume and number of molecules constant, then we can define the temperature as

$$\left. \frac{\partial S}{\partial E} \right|_{V,N} = \frac{1}{T}$$

and we obtain

$$\frac{P(E_s^{(1)})}{P(E_s^{(2)})} = \frac{e^{-E_s^{(1)}/k_B T}}{e^{-E_s^{(2)}/k_B T}}.$$

This implies that

$$P(E_s^{(q)}) \propto e^{-E_s^{(q)}/(k_B T)}$$

and hence the probability of being in a microstate  $q$  is given by

$$P(q) = \frac{1}{Z} e^{-E_q/(k_B T)}, \quad (4.2)$$

where we have written  $E_q$  for the energy of the microstate and  $Z$  is a normalizing factor, known as the *partition function*, defined by

$$Z = \sum_{q \in \mathcal{Q}} e^{-E_q/(k_B T)}.$$

By keeping track of those microstates that correspond to a given system state (also called a macrostate), we can compute the overall probability that a given macrostate is reached.

In order to determine the energy levels associated with different microstates, we will often make use of the *free energy* of the system. Consider an elementary reaction  $A + B \rightleftharpoons AB$ . Let  $E$  be the energy of the system, taken to be operating at pressure  $P$  in a volume  $V$ . The *enthalpy* of the system is defined as  $H = E + PV$  and the *Gibbs free energy* is defined as  $G = H - TS$  where  $T$  is the temperature of the system and  $S$  is its entropy (defined above). The change in bond energy due to the reaction is given by

$$\Delta H = \Delta G + T \Delta S,$$

where the  $\Delta$  represents the change in the respective quantity.  $-\Delta H$  represents the amount of heat that is absorbed from the reservoir, which then affects the entropy of the reservoir.

The resulting formula for the probability of being in a microstate  $q$  is given by

$$P(q) = \frac{1}{Z} e^{-\Delta G/k_B T}.$$

**Example 4.1** (Ligand-receptor binding). To illustrate how these ideas can be applied in a cellular setting, consider the problem of determining the probability that a ligand binds to a receptor protein, as illustrated in Figure 4.1. We model the system by breaking up the cell into  $\Omega$  different locations, each of the size of a ligand molecule, and keeping track of the locations of the  $L$  ligand molecules. The microstates of the system consist of all possible locations of the ligand molecules, including those in which one of the ligand molecules is bound to the receptor molecule.

To compute the probability that the ligand is bound to the receptor, we must compute the energy associated with each possible microstate and then compute the weighted sum of the microstates corresponding to the ligand being bound, normalized by the partition function. We let  $E_{\text{sol}}$  represent the free energy associated with a ligand in free solution and  $E_{\text{bound}}$  represent the free energy associated with the ligand being bound to the receptor. Thus, the energy associated with microstates in which the ligand is not bound to the receptor is given by

$$\Delta G_{\text{sol}} = L E_{\text{sol}}$$

and the energy associated with microstates in which one ligand is bound to the receptor is given by

$$\Delta G_{\text{bound}} = (L - 1) E_{\text{sol}} + E_{\text{bound}}.$$

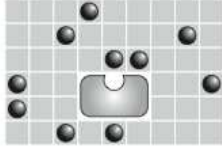
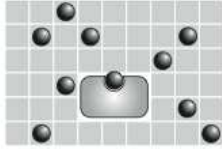
STATE	ENERGY	MULTIPLICITY	WEIGHT
	$L\epsilon_{\text{sol}}$	$\frac{\Omega!}{L!(\Omega-L)!} \approx \frac{\Omega^L}{L!}$	$\frac{\Omega^L}{L!} e^{-\beta L\epsilon_{\text{sol}}}$
	$(L-1)\epsilon_{\text{sol}} + \epsilon_{\text{b}}$	$\frac{\Omega!}{(L-1)!(\Omega-L+1)!} \approx \frac{\Omega^{L-1}}{(L-1)!}$	$\frac{\Omega^{L-1}}{(L-1)!} e^{-\beta[(L-1)\epsilon_{\text{sol}} + \epsilon_{\text{b}}]}$

Figure 4.1: Statistical physics description of ligand-receptor binding. The cell is modeled as a compartment with  $\Omega$  sites, one of which contains a receptor protein. Ligand molecules can occupy any of the sites (first column) and we can compute the Gibbs free energy associated with each configuration (second column). The first row represents all possible microstates in which the receptor protein is not bound, while the second represents all configurations in which one of the ligands binds to the receptor. By accounting for the multiplicity of each microstate (third column), we can compute the weight of the given collection of microstates (fourth column). Figure from Phillips, Kondev and Theriot [56].

Next, we compute the number of possible ways in which each of these two situations can occur. For the unbound ligand, we have  $L$  molecules that can be in any one of  $\Omega$  locations, and hence the total number of combinations is given by

$$N_{\text{sol}} = \binom{\Omega}{L} = \frac{\Omega!}{L!(\Omega-L)!} \approx \frac{\Omega^L}{L!},$$

where the final approximation is valid in the case when  $L \ll \Omega$ . Similarly, the number of microstates in which the ligand is bound to the receptor is

$$N_{\text{sol}} = \binom{\Omega}{L-1} = \frac{\Omega!}{(L-1)!(\Omega-L+1)!} \approx \frac{\Omega^{L-1}}{(L-1)!}.$$

Using these two counts, the partition function for the system is given by

$$Z \approx \frac{\Omega^L}{L!} e^{-\frac{L\epsilon_{\text{sol}}}{k_B T}} + \frac{\Omega^{L-1}}{(L-1)!} e^{-\frac{(L-1)\epsilon_{\text{sol}} + \epsilon_{\text{bound}}}{k_B T}}.$$

Finally, we can compute the steady state probability that the ligand is bound by computing the ratio of the weights for the desired states divided by the partition function

$$P_{\text{bound}} = \frac{1}{Z} \cdot \frac{\Omega^{L-1}}{(L-1)!} e^{-\frac{(L-1)\epsilon_{\text{sol}} + \epsilon_{\text{bound}}}{k_B T}}.$$

▽

While the previous example was carried out for the special case of a ligand molecule binding to a receptor protein, in fact this same type of computation can be used to compute the probability that a transcription factor is attached to a piece of DNA or that two freely moving molecules bind to each other. Each of these cases simply comes down to enumerating all possible microstates, computing the energy associated with each, and then computing the ratio of the sum of the weights for the desired states to the complete partition function.

**Example 4.2** (Transcription factor binding). Suppose that we have a transcription factor  $R$  that binds to a specific target region on a DNA strand (such as the promoter region upstream of a gene). We wish to find the probability  $P_{\text{bound}}$  that the transcription factor will be bound to this location as a function of the number of transcription factor molecules  $n_R$  in the system. If the transcription factor is a repressor, for example, knowing  $P_{\text{bound}}(n_R)$  will allow us to calculate the likelihood of transcription occurring.

To compute the probability of binding, we assume that the transcription factor can bind non-specifically to other sections of the DNA (or other locations in the cell) and we let  $N_{\text{ns}}$  represent the number of such sites. We let  $E_{\text{bound}}$  represent the free energy associated with  $R$  bound to its specified target region and  $E_{\text{ns}}$  represent the free energy for  $R$  in any other non-specific location, where we assume that  $E_{\text{extbound}} < E_{\text{ns}}$ . The microstates of the system consist of all possible assignments of the  $n_R$  transcription factors to either a non-specific location or the target region of the DNA. Since there is only one target site, there can be at most one transcription factor attached there and hence we must count all of the ways in which either zero or one molecule of  $R$  are attached to the target site.

If none of the  $n_R$  copies of  $R$  are bound to the target region then these must be distributed between the  $N_{\text{ns}}$  non-specific locations. Each bound protein has energy  $E_{\text{ns}}$ , so the total energy for any such configuration is  $n_R E_{\text{ns}}$ . The number of such combinations is  $\binom{N_{\text{ns}}}{n_R}$  and so the contribution to the partition function from these microstates is

$$Z_{\text{ns}} = \binom{N_{\text{ns}}}{n_R} e^{-n_R E_{\text{ns}}/(k_B T)} = \frac{N_{\text{ns}}!}{n_R! (N_{\text{ns}} - n_R)!} e^{-n_R E_{\text{ns}}/(k_B T)}$$

For the microstates in which one molecule of  $R$  is bound at a target site and the other  $n_R - 1$  molecules are at the non-specific locations, we have a total energy of  $E_{\text{bound}} + (n_R - 1)E_{\text{ns}}$  and  $\binom{N_{\text{ns}}}{(n_R - 1)}$  possible such states. The resulting contribution to the partition function is

$$Z_{\text{bound}} = \frac{N_{\text{ns}}!}{(n_R - 1)! (N_{\text{ns}} - n_R + 1)!} e^{-(E_{\text{bound}} - (n_R - 1)E_{\text{ns}})/(k_B T)}.$$

The probability that the target site is occupied is now computed by looking at the ratio of the  $Z_{\text{bound}}$  to  $Z = Z_{\text{ns}} + Z_{\text{bound}}$ . After some basic algebraic manipulations,

it can be shown that

$$P_{\text{bound}}(n_R) = \frac{\left(\frac{n_R}{N_{\text{ns}} - n_R + 1}\right) \exp[-(E_{\text{bound}} + E_{\text{ns}})/(k_B T)]}{1 + \left(\frac{n_R}{N_{\text{ns}} - n_R + 1}\right) \exp[-(E_{\text{bound}} + E_{\text{ns}})/(k_B T)]}.$$

If we assume that  $N_{\text{ns}} \gg n_R$ , then we can write

$$P_{\text{bound}}(n_R) \approx \frac{kn_R}{1 + kn_R}, \quad \text{where} \quad k = \frac{1}{N_{\text{ns}}} \exp[-(E_{\text{bound}} - E_{\text{ns}})/(k_B T)].$$

As we would expect, this says that for very small numbers of repressors,  $P_{\text{bound}}$  is close to zero, while for large numbers of repressors,  $P_{\text{bound}} \rightarrow 1$ . The point at which we get a binding probability of 0.5 is when  $n_R = 1/k$ , which depends on the relative binding energies and the number of non-specific binding sites.  $\nabla$

### Chemical Master Equation (CME)

The statistical physics model we have just considered gives a description of the *steady state* properties of the system. In many cases, it is clear that the system reaches this steady state quickly and hence we can reason about the behavior of the system just by modeling the free energy of the system. In other situations, however, we care about the transient behavior of a system or the dynamics of a system that does not have an equilibrium configuration. In these instances, we must extend our formulation to keep track of how quickly the system transitions from one microstate to another, known as the *chemical kinetics* of the system.

To model these dynamics, we return to our enumeration of all possible microstates of the system. Let  $P(q, t)$  represent the probability that the system is in microstate  $q$  at a given time  $t$ . Here  $q$  can be any of the very large number of possible microstates for the system. We wish to write an explicit expression for how  $P(q, t)$  varies as a function of time, from which we can study the stochastic dynamics of the system.

We begin by assuming we have a set of  $M$  reactions  $R_j$ ,  $j = 1, \dots, M$ , with  $\xi_j$  representing the change in state associated with reaction  $R_j$ . The *propensity function* defines the probability that a given reaction occurs in a sufficiently small time step  $dt$ :

$$a_j(q, t)dt = \text{Probability that reaction } R_j \text{ will occur between time } t \text{ and time } t + dt \text{ given that } X(t) = q.$$

The linear dependence on  $dt$  relies on the fact that  $dt$  is chosen sufficiently small. We will typically assume that  $a_j$  does not depend on the time  $t$  and write  $a_j(q)dt$  for the probability that reaction  $j$  occurs in state  $x$ .

Using the propensity function, we can compute the distribution of states at time  $t + dt$  given the distribution at time  $t$ :

$$\begin{aligned} P(q, t + dt | q_0, t_0) &= P(q, t | q_0, t_0) \left( 1 - \sum_{j=1}^M a_j(q) dt \right) + \sum_{j=1}^M P(q - \xi_j | q_0, t_0) a_j(q - \xi_j) dt \\ &= P(q, t | q_0, t_0) + \sum_{j=1}^M \left( a_j(q - \xi_j) P(q - \xi_j, t | q_0, t_0) - a_j(q) P(q, t | q_0, t_0) \right) dt. \end{aligned} \quad (4.3)$$

Since  $dt$  is small, we can take the limit as  $dt \rightarrow 0$  and we obtain the *chemical master equation* (CME):

$$\frac{\partial P}{\partial t}(q, t | q_0, t_0) = \sum_{j=1}^M \left( a_j(q - \xi_j) P(q - \xi_j, t | q_0, t_0) - a_j(q) P(q, t | q_0, t_0) \right) \quad (4.4)$$

This equation is also referred to as the *forward Kolmogorov equation* for a discrete state, continuous time random process.

We will sometimes find it convenient to use a slightly different notation in which we let  $\xi$  represent any transition in the system state (without enumerating the reactions). In this case, we write the propensity function as  $a(\xi; q, t)$ , which represents the incremental probability that we will transition from state  $q$  to state  $q + \xi$  at time  $t$ . When the propensities are not explicitly dependent on time, we simply write  $a(\xi; q)$ . In this notation, the chemical master equation becomes

$$\frac{\partial P}{\partial t}(q, t | q_0, t_0) = \sum_{\xi} \left( a(\xi; q - \xi_j) P(q - \xi_j, t | q_0, t_0) - a(\xi; q) P(q, t | q_0, t_0) \right), \quad (4.5)$$

where the sum is understood to be over all allowable transitions.

Under some additional assumptions, we can rewrite the master equation in differential form as

$$\frac{d}{dt} P(q, t) = \sum_{\xi} a(\xi; q - \xi) P(q - \xi, t) - \sum_{\xi} a(\xi; q) P(q, t), \quad (4.6)$$

where we have dropped the dependence on the initial condition for notational convenience. We see that the master equation is a *linear* differential equation with state  $P(q, t)$ . However, it is important to note that the size of the state vector can be very large: we must keep track of the probability of every possible microstate of the system. For example, in the case of the ligand-receptor problem discussed earlier, this has a factorial number of states based on the number of possible sites in the model. Hence, even for very simple systems, the master equation cannot typically be solved either analytically or in a numerically efficient fashion.

Despite its complexity, the master equation does capture many of the important details of the chemical physics of the system and we shall use it as our basic representation of the underlying dynamics. As we shall see, starting from this equation we can then derive a variety of alternative approximations that allow us to answer specific equations of interest.

The key element of the master equation is the propensity function  $a(\xi; q, t)$ , which governs the rate of transition between microstates. Although the detailed value of the propensity function can be quite complex, its functional form is often relatively simple. In particular, for a unimolecular reaction  $\xi$  of the form  $A \rightarrow B$ , the propensity function is proportional to the number of molecules of A that are present:

$$a(\xi; q, t) = c_{\xi} n_A. \quad (4.7)$$

This follows from the fact that each reaction is independent and hence the likelihood of a reaction happening depends directly on the number of copies of A that are present.

Similarly, for a bimolecular reaction, we have that the likelihood of a reaction occurring is proportional to the product of the number of molecules of each type that are present (since this is the number of independent reactions that can occur). Hence, for a reaction  $\xi$  of the form  $A + B \rightarrow C$  we have

$$a(\xi; q, t) = c_{\xi} n_A n_B. \quad (4.8)$$

The rigorous verification of this functional form is beyond the scope of this text, but roughly we keep track of the likelihood of a single reaction occurring between A and B and then multiply by the total number of combinations of the two molecules that can react ( $n_A \cdot n_B$ ).

A special case of a bimolecular reaction occurs when  $A = B$ , so that our reaction is given by  $2A \rightarrow B$ . In this case we must take into account that a molecule cannot react with itself, and so the propensity function is of the form

$$a(\xi; q, t) = c_{\xi} n_A (n_A - 1). \quad (4.9)$$

Although it is tempting to extend this formula to the case of more than two species being involved in a reaction, usually such reactions actually involve combinations of bimolecular reactions, e.g.:



This more detailed description reflects that fact that it is extremely unlikely that three molecules will all come together at precisely the same instant, versus the much more likely possibility that two molecules will initially react, followed by a second reaction involving the third molecule.

The propensity functions for these cases and some others are given in Table 4.1.

Table 4.1: Examples of propensity functions for some common cases [29]. Here we take  $r_a$  and  $r_b$  to be the effective radii of the molecules,  $m^* = m_a m_b / (m_a + m_b)$  is the reduced mass of the two molecules,  $\Omega$  is the volume over which the reaction occurs,  $T$  is temperature,  $k_B$  is Boltzmann's constant and  $n_a, n_b$  are the numbers of molecules of  $A$  and  $B$  present.

Reaction type	Propensity function coefficient, $c_\xi$
Reaction occurs if molecules "touch"	$\Omega^{-1} \left( \frac{8k_B T}{\pi m^*} \right)^{1/2} \pi (r_a + r_b)^2$
Reaction occurs if molecules collide with energy $\epsilon$	$\Omega^{-1} \left( \frac{8k_B T}{\pi m^*} \right)^{1/2} \pi (r_a + r_b)^2 \cdot e^{-\epsilon/k_B T}$
Steady state transcription factor	$P_{\text{bound}} k_{\text{oc}} n_{\text{RNAP}}$

**Example 4.3** (Transcription of mRNA). Consider the production of mRNA from a single copy of DNA. We have two basic reactions that can occur: mRNA can be produced by RNA polymerase transcribing the DNA and producing an mRNA strand, or mRNA can be degraded. We represent the microstate  $q$  of the system in terms of the number of mRNA's that are present, which we write as  $n$  for ease of notation. The reactions can now be represented as  $\xi = +1$ , corresponding to transcription and  $\xi = -1$ , corresponding to degradation. We choose as our propensity functions

$$a(+1; n, t) = \alpha, \quad a(-1; n, t) = \gamma n,$$

by which we mean that the probability of that a gene is transcribed in time  $dt$  is  $\alpha dt$  and the probability that a transcript in time  $dt$  is  $\gamma n dt$  (proportional to the number of mRNA's).

We can now write down the master equation as described above. Equation (4.3) becomes

$$\begin{aligned} P(n, t + dt) &= P(n, t) \left( 1 - \sum_{\xi=+1, -1} a(\xi; n, t) dt \right) + \sum_{\xi=+1, -1} P(n - \xi, t) a(\xi; n - \xi, t) dt \\ &= P(n, t) - a(+1; n, t) P(n, t) - a(-1; n, t) P(n, t) \\ &\quad + a(+1; n - 1, t) P(n - 1, t) + a(-1; n + 1, t) P(n + 1) \\ &= P(n, t) + \alpha P(n - 1, t) dt - (\alpha + \gamma n) P(n, t) dt + \gamma(n + 1) P(n + 1, t) dt. \end{aligned}$$

This formula holds for  $n > 0$ , with the  $n = 0$  case satisfying

$$P(0, t + dt) = P(0, t) - \alpha P(0, t) dt + \gamma P(1, t) dt.$$

Notice that we have an infinite number of equations, since  $n$  can be any positive integer.

We can write the differential equation version of the master equation by subtracting the first term on the right hand side and dividing by  $dt$ :

$$\begin{aligned} \frac{d}{dt} P(n, t) &= \alpha P(n - 1, t) - (\alpha + \gamma n) P(n, t) + \gamma(n + 1) P(n + 1, t), \quad n > 0 \\ \frac{d}{dt} P(0, t) &= -\alpha P(0, t) + \gamma P(1, t). \end{aligned}$$

Again, this is an infinite number of differential equations, although we could take some limit  $N$  and simply declare that  $P(N, t) = 0$  to yield a finite number.

One simple type of analysis that can be done on this equation without truncating it to a finite number is to look for a steady state solution to the equation. In this case, we set  $\dot{P}(n, t) = 0$  and look for a constant solution  $P(n, t) = p_e(n)$ . This yields an algebraic set of relations

$$\begin{aligned} 0 &= -\alpha p_e(0) + \gamma p_e(1) & \implies & \alpha p_e(0) = \gamma p_e(1) \\ 0 &= \alpha p_e(0) - (\alpha + \gamma) p_e(1) + 2\gamma p_e(2) & & \alpha p_e(1) = 2\gamma p_e(2) \\ 0 &= \alpha p_e(1) - (\alpha + 2\gamma) p_e(2) + 3\gamma p_e(3) & & \alpha p_e(2) = 3\gamma p_e(3) \\ & \vdots & & \vdots \\ & & & \alpha p_e(n-1) = n\gamma p_e(n). \end{aligned}$$

It follows that the distribution of steady state probabilities is given by the Poisson distribution

$$p(n) = e^{-\alpha/\gamma} \frac{(\alpha/\gamma)^n}{n!},$$

and the mean, variance and coefficient of variation are thus

$$\mu = \frac{\alpha}{\gamma}, \quad \sigma^2 = \frac{\alpha}{\gamma}, \quad CV = \frac{\mu}{\sigma} = \frac{1}{\sqrt{\mu}} = \sqrt{\frac{\gamma}{\alpha}}.$$

▽

### Chemical Langevin equation (CLE)

The chemical master equation gives a complete description of the evolution of the distribution of a system, but it can often be quite cumbersome to work with directly. A number of approximations to the master equation are thus used to provide more tractable formulations of the dynamics. The first of these that we shall consider is known as the *chemical Langevin equation* (CLE).

To derive the chemical Langevin equation, we start by assuming that the number of species in the system is large and that we can therefore represent the system using a vector of real numbers  $X$ , with  $X_i$  representing the (real-valued) number of molecules in  $S_i$ . (Often  $X_i$  will be divided by the volume to give a real-valued concentration of species  $S_i$ .) In addition, we assume that we are interested in the dynamics on time scales in which individual reactions are not important and so we can look at how the system state changes over time intervals in which many reactions occur and hence the system state evolves in a smooth fashion.

Let  $X(t)$  be the state vector for the system, where we assume now that the elements of  $X$  are real-valued rather than integer valued. We make the further approximation that we can lump together multiple reactions so that instead of keeping track of the individual reactions, we can average across a number of reactions over

a time  $\tau$  to allow the continuous state to evolve in continuous time. The resulting dynamics can be described by a stochastic process of the form

$$X_i(t + \tau) = X_i(t) + \sum_{j=1}^M \xi_{ij} a_j(X(t)) \tau + \sum_{j=1}^M \xi_{ij} a_j^{1/2}(X(t)) \mathcal{N}_j(0, \sqrt{\tau}),$$

where  $a_j$  are the propensity functions for the individual reactions,  $\xi_{ij}$  are the corresponding changes in the system states  $X_i$  and  $\mathcal{N}_j$  are a set of independent Gaussian random variables with zero mean and variance  $\tau$ .

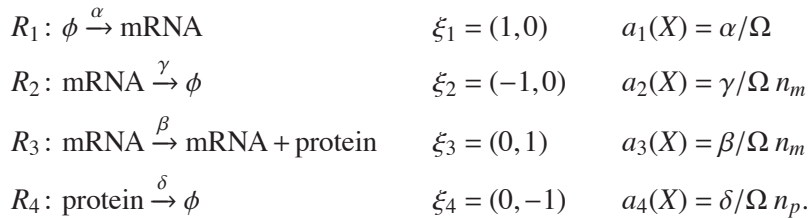
If we assume that  $\tau$  is small enough that we can use the derivative to approximate the previous equation (but still large enough that we can average over multiple reactions), then we can write

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^M \xi_{ji} a_j(X(t)) + \sum_{j=1}^M \xi_{ji} a_j^{1/2}(X(t)) \Gamma_j(t) =: A_i(X(t)) + \sum_{j=1}^M B_{ij}(X(t)) \Gamma_j(t), \quad (4.10)$$

where  $\Gamma_j$  are white noise processes. This equation is called the *chemical Langevin equation* (CLE).

**Example 4.4** (Protein production). Consider a simplified model of protein production in which mRNAs are produced by transcription and proteins by translation. We also include degradation of both mRNAs and proteins, but we do not model the detailed processes of elongation of the mRNA and polypeptide chains.

We can capture the state of the system by keeping track of the number of copies of mRNA and proteins. We further approximate this by assuming that the number of each of these is sufficiently large that we can keep track of its concentration, and hence  $X = (n_m, n_p)$  where  $n_m \in \mathbb{R}$  is the amount of mRNA and  $n_p \in \mathbb{R}$  is the concentration of protein. Letting  $\Omega$  represent the volume, the reactions that govern the dynamics of the system are given by:



Substituting these expressions into equation (4.10), we obtain a stochastic differential equation of the form

$$\frac{d}{dt} \begin{pmatrix} n_m \\ n_p \end{pmatrix} = \begin{pmatrix} -\gamma/\Omega & 0 \\ \beta/\Omega & -\delta/\Omega \end{pmatrix} \begin{pmatrix} n_m \\ n_p \end{pmatrix} + \begin{pmatrix} \alpha/\Omega \\ 0 \end{pmatrix} + \begin{pmatrix} (\sqrt{\alpha/\Omega} + \sqrt{\gamma n_m/\Omega}) \Gamma_m \\ (\sqrt{\beta n_m/\Omega} + \sqrt{\delta n_p/\Omega}) \Gamma_p \end{pmatrix},$$

where  $\Gamma_m$  and  $\Gamma_p$  are independent white noise processes with unit variance. (Note that in deriving this equation we have used the fact that the sum of two independent Gaussian processes is a Gaussian process.) ▽

### Fokker-Planck equations (FPE)

The chemical Langevin equation provides a stochastic ordinary differential equation that describes the evolution of the system state. A slightly different (but completely equivalent) representation of the dynamics is to model how the probability distribution  $P(q, t)$  evolves in time. As in the case of the chemical Langevin equation, we will assume that the system state is continuous and write down a formula for the evolution of the density function  $p(x, t)$ . This formula is known as the *Fokker-Planck equations* (FPE) and is essentially an approximation on the chemical master equation.

Consider first the case of a random process in one dimension. We assume that the random process is in the same form as the previous section:

$$\frac{dX(t)}{dt} = A(X(t)) + B(X(t))\Gamma(t). \quad (4.11)$$

The function  $A(X)$  is called the *drift term* and  $B(X)$  is the *diffusion term*. It can be shown that the probability density function for  $X$ ,  $p(x, t | x_0, t_0)$ , satisfies the partial differential equation

$$\frac{\partial p}{\partial t}(x, t | x_0, t_0) = -\frac{\partial}{\partial x}(A(x, t)p(x, t | x_0, t_0)) + \frac{1}{2} \frac{\partial^2}{\partial x^2}(B^2(x, t)p(x, t | x_0, t_0)) \quad (4.12)$$

Note that here we have shifted to the probability density function since we are considering  $X$  to be a continuous state random process.

In the multivariate case, a bit more care is required. Using the chemical Langevin equation (4.10), we define

$$D_i(x, t) = \sum_{j=1}^M B_{ij}^2(x, t), \quad C_{ij}(x, t) = \sum_{k=1}^M B_{ik}(x, t)B_{jk}(x, t), \quad i < j = 1, \dots, M.$$

The Fokker-Planck equation now becomes

$$\begin{aligned} \frac{\partial p}{\partial t}(x, t | x_0, t_0) = & -\sum_{i=1}^M \frac{\partial}{\partial x_i}(A_i(x, t)p(x, t | x_0, t_0)) \\ & + \frac{1}{2} \sum_{i=1}^M \frac{\partial}{\partial x_i} \frac{\partial^2}{\partial x_i^2}(D_i(x, t)p(x, t | x_0, t_0)) \\ & + \sum_{\substack{i, j=1 \\ i < j}}^M \frac{\partial^2}{\partial x_i \partial x_j}(C_{ij}(x, t)p(x, t | x_0, t_0)). \end{aligned} \quad (4.13)$$

### Linear noise approximation (LNA)

The chemical Langevin equation and the Fokker-Planck equation provide approximations to the chemical master equation. A slightly different approximation can be obtained by expanding the density function in terms of a size parameter  $\Omega$ . This approximation is known as the *linear noise approximation* (LNA) or the  $\Omega$  *expansion* [43].

We begin with a master equation for a continuous random variable  $X$ , which we take to be of the form

$$\frac{\partial p}{\partial t}(x, t) = \int (a_{\Omega}(\xi; x - \xi)p(x - \xi, t) - a_{\Omega}(\xi; x)p(x, t)) d\xi,$$

where we have dropped the dependence on the initial condition for notational simplicity. As before, the propensity function  $a_{\Omega}(\xi; x)$  represents the transition probability between a state  $x$  and a state  $x + \xi$  and we assume that it is a function of a parameter  $\Omega$  that represents the size of the system (typically the volume). Since we are working with continuous variables, we now have an integral in place of our previous sum.

We assume that the mean of  $X$  can be written as  $\Omega\phi(t)$  where  $\phi(t)$  is a continuous function of time that represents the evolution of the mean of  $X/\Omega$ . To understand the fluctuations of the system about this mean, we write

$$X = \Omega\phi + \Omega^{\frac{1}{2}}Z,$$

where  $Z$  is a new variable representing the perturbations of the system about its mean. We can write the distribution for  $Z$  as

$$p_Z(z, t) = p_X(\Omega\phi(t) + \Omega^{\frac{1}{2}}z, t)$$

and it follows that the derivatives of  $p_Z$  can be written as

$$\begin{aligned} \frac{\partial^y p_Z}{z^y} &= \Omega^{\frac{1}{2}y} \frac{\partial^y p_X}{x^y} \\ \frac{\partial p_Z}{\partial t} &= \frac{\partial p_X}{\partial t} + \Omega \frac{d\phi}{dt} \frac{\partial p_X}{\partial x} = \frac{\partial p_X}{\partial t} + \Omega^{\frac{1}{2}} \frac{d\phi}{dt} \frac{\partial p_Z}{\partial z}. \end{aligned}$$

We further assume that the  $\Omega$  dependence of the propensity function is such that

$$a_{\Omega}(\xi, \Omega\phi) = f(\Omega)\tilde{a}(\xi; \phi),$$

where  $\tilde{a}$  is not dependent on  $\Omega$ . From these relations, we can now derive the master equation for  $p_Z$  in terms of powers of  $\Omega$  (derivation omitted).

The  $\Omega^{1/2}$  term in the expansion turns out to yield

$$\frac{d\phi}{dt} = \int \xi a(\xi, \Omega\phi) d\xi, \quad \phi(0) = \frac{X(0)}{\Omega},$$

which is precisely the equation for the mean of the concentration. It can further be shown that the terms in  $\Omega^0$  are given by

$$\frac{\partial p_Z(z, \tau)}{\partial \tau} = -\alpha'_1(\phi) \frac{\partial}{\partial z} (z p_Z(z, t)) + \frac{1}{2} \alpha_2(\phi) \frac{\partial^2 p_Z(z, t)}{\partial z^2}, \quad (4.14)$$

where

$$\alpha_\nu(x) = \int \xi^\nu \tilde{a}(\xi; x) d\xi, \quad \tau = \Omega^{-1} f(\Omega) t.$$

Notice that in the case that  $\phi(t) = \phi_0$ , this equation becomes the Fokker-Planck equation derived previously.

Higher order approximations to this equation can also be carried out by keeping track of the expansion terms in higher order powers of  $\Omega$ . In the case where  $\Omega$  represents the volume of the system, the next term in the expansion is  $\Omega^{-1}$  and this represents fluctuations that are on the order of a single molecule, which can usually be ignored.

### Rate reaction equations (RRE)

As we already saw in Chapter 2, the reaction rate equations can be used to describe the dynamics of a chemical system in the case where there are a large number of molecules whose state can be approximated using just the concentrations of the molecules. We re-derive the results from Section 2.1 here, being more careful to point out what approximations are being made.

We start with the chemical Langevin equations (4.10), from which we can write the dynamics for the average quantity of the each species at each point in time:

$$\frac{d\langle X_i(t) \rangle}{dt} = \sum_{j=1}^M \xi_{ji} \langle a_j(X(t)) \rangle,$$

where the second order term drops out under the assumption that the  $\Gamma_j$ 's are independent processes. We see that the reaction rate equations follow by defining  $x_i = \langle X_i \rangle / \Omega$  and *assuming* that  $\langle a_j(X(t)) \rangle = a_j(\langle X(t) \rangle)$ . This relationship is true when  $a_j$  is linear (e.g., in the case of a unimolecular reaction), but is an approximation otherwise.

**4.2 Simulation of Stochastic sections****4.3 Analysis of Stochastic Systems****4.4 Linearized Modeling and Analysis****4.5 Markov chain modeling and analysis****4.6 System identification techniques****4.7 Model Reduction****Exercises**

**4.1** Consider gene expression:  $\phi \xrightarrow{k} m$ ,  $m \xrightarrow{\beta} m + P$ ,  $m \xrightarrow{\gamma} \phi$ , and  $P \xrightarrow{\delta} \phi$ . Answer the following questions:

(a) Use the stochastic simulation algorithm (SSA) to obtain realizations of the stochastic process of gene expression and numerically compare with the deterministic ODE solution. Explore how the realizations become close to or apart from the ODE solution when the volume is changed. Determine the stationary probability distribution for the protein (you can do this numerically, but note that this process is linear, so you can compute the probability distribution analytically in closed form).

(b) Now consider the additional binding reaction of protein P with downstream DNA binding sites D:  $P + D \xrightleftharpoons[k_{off}]{k_{on}} C$ . Note that the system no longer linear due to the presence of a bi-molecular reaction. Use the SSA algorithm to obtain sample realizations and numerically compute the probability distribution of the protein and compare it to what you obtained in part (a). Explore how this probability distribution and the one of C change as the rates  $k_{on}$  and  $k_{off}$  become larger and larger with respect to  $\delta, k, \beta, \gamma$ . Do you think we can use a QSS approximation similar to what we have done for ODE models?

(c) Determine the Langevin equation for the system in part (b) and obtain sample realizations. Explore numerically how good this approximation is when the volume decreases/increases.

