2

Dynamical Representations of Molecular Networks

A living cell can be viewed as a huge dynamical system or molecular network with stochastic fluctuations at the molecular levels, where cellular components interact dynamically, both temporally and spatially. Dynamical representations of the molecular network are necessary for an accurate understanding of the temporal and spatial evolution of a cellular system and can also provide deep insight into the dynamical interactions among the cellular components. This chapter describes a general theoretical framework to model molecular networks with the consideration of discrete state transitions and stochastic fluctuations. In particular, we provide both stochastic and deterministic formulation to model networks of biochemical reactions in a cell.

2.1 Biochemical Reactions

Biochemical reactions, which specify how the states of a system change and how fast that change occurs, are very important in modeling molecular networks. Many complex networks can be expressed by such reactions, either qualitatively or quantitatively. Biochemical reactions are so fundamental that the same list of biochemical reactions can lead to different models, *e.g.*, a graphical representation, a stochastic process, or a deterministic model. In this sense, representing processes by biochemical reactions is more basic than using a mathematical model of either stochastic or deterministic dynamics.

Writing down a list of biochemical reactions corresponding to a cellular system or a molecular network, together with the rate of every reaction and the initial condition of each species, is a powerful and flexible way to specify the system. However, the reactions themselves specify only the qualitative structure of the system and must be augmented with additional tools or formalisms before they can be used to analyze and simulate dynamics of the network and further make predictions. One tool is the dynamical representations of cellular systems or molecular networks.

In this section, we introduce some basic biochemical reactions involved in cellular systems. A cellular system consists of a network of coupled biochemical reactions. These reactions can be transcription, translation, dimerization, protein or mRNA degradation, enzyme-catalyzed reactions, transportation, diffusion, binding or unbinding, DNA or histone methylation, histone acetylation or phosphorylation. These biochemical reactions constitute various biomolecular networks, *e.g.*, metabolic, genetic, and signaling networks.

In an elementary biochemical reaction, one or more biochemical species react directly to form products in a single reaction step with a single transition state. More complicated reactions can be decomposed into a sequence of elementary reactions. A general elementary biochemical reaction can be represented as follows:

$$r_1R_1 + r_2R_2 + \cdots + r_mR_m \xrightarrow{k} p_1P_1 + p_2P_2 + \cdots + p_nP_n, \tag{2.1}$$

where m is the number of reactants, and n is the number of products. The terms to the left of the arrow are called reactants, and those on the right are called products. Thus, R_i is the ith reactant, and P_j is the jth product. r_i and p_j are the numbers of reactant R_i consumed and product P_j produced in a reaction step, respectively. k is a positive number to represent the rate of the reaction. The coefficients r_i and p_j are known as stoichiometries and are generally small positive integers. The reactants and products in a cell are DNA, RNA, proteins, or other chemicals. The total reaction order is defined as the sum of r_i, *i.e.*, $\sum_{i=1}^{m} r_i$.

The general reaction (2.1) can express any biochemical reaction in a cell. However, in order to clearly represent concrete biochemical processes, we sequentially describe typical reactions in cellular systems. First, we consider the dimerization of a protein P with reaction rate k_d:

$$2P \xrightarrow{k_d} P_2, \tag{2.2}$$

where $2P$ means $P + P$. (2.2) is a reaction for a protein–protein interaction, which results in a homodimer P_2 composed of two protein monomers P. The reaction has one reactant P and one product P_2 with stoichiometries of 2 and 1, respectively. Stoichiometries of 1 are not usually written explicitly. Similarly, the reaction for the dissociation of the dimer P_2 with reaction rate k_{-d} is written as follows:

$$P_2 \xrightarrow{k_{-d}} 2P. \tag{2.3}$$

A reaction that can happen in both directions is termed reversible. Reversible reactions are quite common in biological systems and can be briefly written by one reaction equation. For instance, (2.2)–(2.3) can be represented as

$$2P \underset{k_{-d}}{\overset{k_d}{\rightleftharpoons}} P_2 \tag{2.4}$$

or sometimes as

$$2P \overset{K_{eq}}{\rightleftharpoons} P_2. \tag{2.5}$$

Here, $K_{eq} = k_d/k_{-d}$ is called the equilibrium constant.

It is important to remember that the notation for a reversible reaction is simply a convenient way to represent two separate reactions. When modeling, we still need to decompose a reversible reaction or more complex reactions into a sequence of elementary reactions in order to analyze and simulate their dynamics. An elementary reaction is a biochemical reaction in which one or more of the chemical species react directly to form products in a single reaction step and with a single transition state.

Before introducing different modeling approaches to systems of coupled biochemical reactions in the next section, we will first describe in detail some basic biochemical processes and show how their essential features can be captured with fairly simple systems of coupled biochemical reactions. Generally, a complex molecular network is intertwined among various processes, including gene regulation, protein interactions, metabolism, and signal transduction. Reactions in each process are relatively strongly correlated, but reactions between these processes are relatively independent, and therefore we only need to couple all the reactions involved in the system or process of interest when modeling it.

Transcription is a key cellular process for gene regulation, and control of transcription is a fundamental regulation mechanism in a living organism. The crucial stage of the transcription process is the binding of a RNAP to the promoter of a gene, which is regulated by various TFs, to initiate the transcription or produce mRNA. A RNAP is an enzyme which copies the genetic sequence of a gene and synthesizes the mRNA by attaching to the DNA strand. A TF or a sequence-specific DNA binding factor is a protein that binds to specific DNA sequences and thereby controls the transcription of genetic information from DNA to RNA by promoting or blocking the recruitment of RNAP to specific genes as an activator or a repressor. Note that a particular feature of TFs is that they contain one or more DNA binding domains which attach to specific DNA sequences adjacent to the genes that they regulate. Other proteins or chemicals without DNA binding domains also play crucial roles in gene regulation, *e.g.*, by binding to a TF to form a transcriptional complex, and are called as cofactors. In this book, we consider a TF as one protein or one complex unless otherwise specified.

Consider the case of one TF P, which can be a monomer or a complex, and one free DNA binding site D in the promoter region of a gene, which produces mRNA. The transcription process as a system of coupled biochemical reactions can be represented as follows:

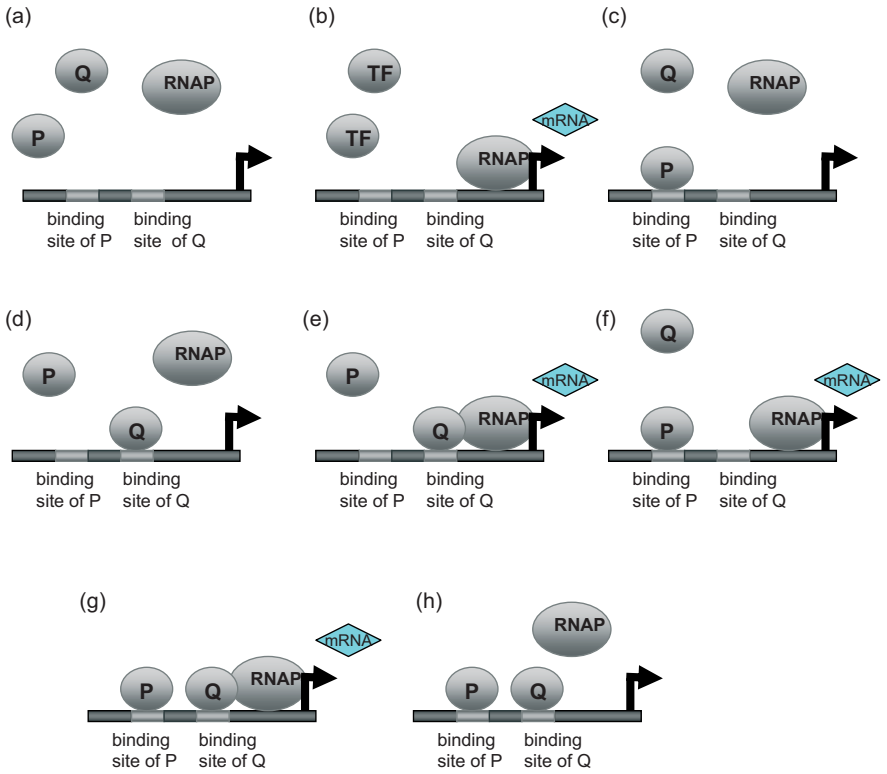

Figure 2.1 Transcription regulation of two TFs (P and Q) and two binding sites on the promoter of a gene. There are eight cases for the transcription regulation: (a) no binding, (b) RNAP binding, (c) P bound to the site, (d) Q bound to the site, (e) Q bound to the site with RNAP binding, (f) P bound to the site with RNAP binding, (g) P and Q bound to the sites with RNAP binding, and (h) P and Q bound to the sites, among which RNAP binds to the promoter to produce mRNA for four cases, *i.e.*, (b), (e), (f), and (g)

$$P + D \underset{k_{-a}}{\overset{k_a}{\rightleftharpoons}} P \cdot D, \tag{2.6}$$

$$D \xrightarrow{k_d} mRNA + D, \tag{2.7}$$

$$P \cdot D \xrightarrow{k_{dc}} mRNA + P \cdot D, \tag{2.8}$$

where k_a, k_{-a}, k_d, and k_{dc} are rate constants for the respective reactions, and $P \cdot D$ is a protein–DNA complex which denotes TF P bound to the binding site D in the promoter. Note that we do not explicitly include RNAP in the model to simplify the expression, although RNAP is necessary in initiating the transcription process. Transcriptional reactions (2.7) and (2.8) indicate

that transcription rates k_d and k_{dc} to produce mRNA are different without TF and with TF. If the TF is an activator to enhance the transcription (*i.e.*, TF recruits RNAP binding to the promoter to initiate the transcription), k_{dc} will be clearly larger than k_d. On the other hand, if the TF is a repressor to inhibit the transcription (*i.e.*, TF prevents RNAP binding to the promoter and from initiating the transcription), k_{dc} will be smaller than k_d. Unlike a single binding reaction (2.6), transcription reactions (2.7) and (2.8) are actually a chain of reactions due to the synthesis of a RNA sequence from a number of nucleotides and are generally slow.

Sometimes, there are multiple TFs which can bind to different binding sites to regulate the expression of the same gene. Such a case can be modeled in a similar way. For example, consider the case where two TFs, P and Q, can bind to two respective binding sites on the DNA, as shown in Figure 2.1. Using similar notation $P \cdot D$ to denote P bound to the binding site D, and $Q \cdot P \cdot D$ to denote Q bound to $P \cdot D$ or P bound to $\mathrm{Q} \cdot \mathrm{D}$, such processes can be formulated by a set of reactions as follows:

$$P + D \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P \cdot D, \tag{2.9}$$

$$Q + D \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} Q \cdot D, \tag{2.10}$$

$$P + Q \cdot D \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} P \cdot Q \cdot D, \tag{2.11}$$

$$Q + P \cdot D \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} P \cdot Q \cdot D, \tag{2.12}$$

$$D \xrightarrow{k_5} mRNA + D, \tag{2.13}$$

$$P \cdot D \xrightarrow{k_6} mRNA + P \cdot D, \tag{2.14}$$

$$Q \cdot D \xrightarrow{k_7} mRNA + Q \cdot D, \tag{2.15}$$

$$P \cdot Q \cdot D \xrightarrow{k_8} mRNA + P \cdot Q \cdot D. \tag{2.16}$$

Clearly, the transcription reactions (2.13), (2.14), (2.15), and (2.16) correspond to cases (b), (f), (e), and (g) of Figure 2.1, respectively.

Transcription of a gene occurs when RNAP is bound to the promoter of DNA regulated by the TF P (or multiple TFs). Thus, the DNA binding site is either free D or bound $P \cdot D$, resulting in a conservation equation

$$D + P \cdot D = n_x \tag{2.17}$$

for the system (2.7)–(2.8), or

$$D + P \cdot D + Q \cdot D + P \cdot Q \cdot D = n_x \tag{2.18}$$

for the system (2.13)–(2.16), where n_x is the total binding sites of the promoter. Note that we also use the same symbols D, P, Q, $Q \cdot D$, $P \cdot D$, and

$P \cdot Q \cdot D$ in (2.17)–(2.18) to represent the concentrations of the respective molecules.

Similarly to the transcription processes, translation is another complicated process which sometimes involves over several hundred reactions to produce a single protein from a single mRNA. The key stages involved in the translation process are the binding of a ribosome to the mRNA, and the translation of the mRNA to a polypeptide chain, and then to a functional protein. For instance, the production of a protein and a complex of multiple proteins can be formulated as a set of reactions as follows:

$$mRNA \xrightarrow{k_p} P + mRNA, \tag{2.19}$$

$$P + P \underset{k_{-b}}{\overset{k_b}{\rightleftharpoons}} P_2, \tag{2.20}$$

$$P_2 + P_2 \underset{k_{-c}}{\overset{k_c}{\rightleftharpoons}} P_4, \tag{2.21}$$

where P, P_2, and P_4 are a protein monomer, a dimer, and a tetramer, respectively. Similarly to the transcription reactions, the translation process (2.19) is also a chain of reactions. Note that we do not model every aspect of the translation and production of protein complex processes, but express the prominent stages of the translation and features of interest. In particular, we do not consider the processes of binding of ribosomes to the mRNA and the folding of the polypeptide chain into a functional protein.

Degradation of mRNA and protein can be modeled in a similar way:

$$mRNA \xrightarrow{d_m} 0, \tag{2.22}$$

$$P \xrightarrow{d_p} 0, \tag{2.23}$$

which mean that mRNA or protein is transformed or degraded into nothing, where d_m and d_p are the degradation rates. In fact, mRNA and protein degradation is a rather complex process. Note that the transcription processes (2.7)–(2.8) and (2.13)–(2.16), translation process (2.19), and degradation processes (2.22)–(2.23) all are a chain of reactions, which are generally much slower than the binding and unbinding processes (2.9)–(2.12) and (2.20)–(2.21).

The transportation process is also important, especially, in eukaryotes. Due to the existence of a nucleus, mRNA must be transported out of the cell nucleus before the translation process occurs. In contrast, proteins (*e.g.*, TFs) must be transported into the nucleus from cytoplasm to regulate gene expression. Transduction has also been shown to be important, especially, in higher eukaryotes and multicellular organisms, because individual cells need to communicate with each other using chemical signaling molecules which can dissolve in the cytosol and diffuse between individual cells and their extracellular medium. By appropriately modeling this process, it might become possible to elucidate the fundamental mechanisms underlying many biological phenomena, *e.g.*, collective behavior through intercellular signaling and

signaling pathways. A model for this process would be as simple as

$$I \xrightarrow{k} A, \tag{2.24}$$

where I denotes an mRNA or a protein within the nucleus, and A corresponds to the mRNA or the protein in the cytoplasm, or *vice versa*, for the transportation process. On the other hand, for the transduction process, I and A denote intracellular and extracellular signaling molecules, respectively. Such a reaction can also be used to model the transition between different states such as inactive and active states of a protein.

Metabolism is a general term for catabolic and anabolic reactions. It is a highly organized process and often involves thousands of reactions that are catalyzed by enzymes. Phosphorylation in signaling cascades is also an enzyme kinetic reaction with the kinase facilitating the phosphorylation of a substrate. Consider the simple enzyme-catalyzed biochemical reactions

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P. \tag{2.25}$$

The reactions comprise a reversible formation of an enzyme–substrate complex ES from the free enzyme E and the substrate S and an irreversible release of the product P. Generally, the duration of the state ES is very short. In the reactions, the enzyme is neither produced nor consumed. Therefore, its total concentration remains constant. It may be free or involved in the complex. Note that in this book, without confusion, we use both XY and $X \cdot Y$ to express the binding or complex of X and Y.

Besides the reactions mentioned above, there are many other reactions involved in cellular systems that we have omitted, such as post-transcriptional regulation, post-translational modification, and microRNA regulation. Although involved in different processes, many reactions can be represented in similar forms. For example, the reaction form for the formation of a heterodimer is the same as that for the formation of an enzyme–substrate complex in the enzyme-catalyzed reaction. Moreover, small RNA-mediated post-transcriptional regulation can be described similarly as the formation of a heterodimer because the small RNA (sRNA) itself is consumed via the interaction between sRNAs and their mRNA targets.

Although there are many kinds of biochemical processes and reactions, after they are decomposed into elementary reactions, most of them can be classified into monomolecular reactions (the first-order reactions), bimolecular reactions (the second-order reactions), or trimolecular reactions (the third-order reactions) according to the number of molecules involved. The monomolecular reaction has the form of degradation (2.22)–(2.23) or transportation and transduction (2.24). When the reactant concentration in monomolecular reactions is considered to be constant, the first-order reactions become the zero-order ones. The bimolecular reactions are probably the most common reactions in cellular systems, *e.g.*, (2.9)–(2.12) and (2.20)–(2.21). There are two different

ways by which two reactants combine to form one product. One is the binding of two same molecules such as the formation of a homodimer, and the other is the binding of two different molecules to form a heterodimer, *e.g.*, the formation of an enzyme–substrate complex, the binding of a TF to a promoter, and the binding of a small RNA to its mRNA target. Trimolecular reactions or higher-order reactions are rare because the probability of the simultaneous collision of three or more molecules is very small. A more general reaction can be represented as (2.1), which is the $\sum_{i=1}^{m} r_i$-order reaction.

2.2 Molecular Networks

Although biochemical reactions are fundamental to understanding a cellular system at the molecular level due to the detailed information that these reactions provide, how these reactions form various molecular networks to facilitate specific cellular functions is unclear. Hence, it is important to elucidate not only the function of each individual reaction but also that of the associated molecular network as a whole. Note that we particularly study biological networks at the molecular level, and therefore, a biological network simply means a biomolecular network or a molecular network in this book. Actually, a living organism can be viewed as a huge biochemical reaction network, which is nonlinear with both intrinsic and extrinsic stochastic fluctuations. A molecular network is assumed to organize a list of biochemical reactions in an accurate, complete, and comprehensive manner. Modeling and analyzing the molecular networks may not only lead to the elucidation of the essential mechanism of how biochemical reactions generate various particular cellular functions but also lead to the revelation of the regulatory roles of individual reactions in the network functions.

Many types of molecular networks, such as gene regulatory networks, transcription regulatory networks, protein interaction networks, metabolic networks, and signal transduction networks exist in a cell (Cao and Liang 2008, Chen *et al.* 2009, Jeong *et al.* 2000, Tyson *et al.* 2001). However, few such networks are known for their complete structures, even in the simplest bacteria. Still less is known on how the networks interact at different levels in a cell, and how to predict the complete state description of a eukaryotic cell or a bacterial organism at a given point in the future. In this sense, the study of molecular networks from a systems biology perspective is still in its infancy. Next, we show how a molecular network is represented by a graph, a stochastic system, a deterministic system or a hybrid system, depending on the requirements to model and analyze a specific cellular system.

2.3 Graphical Representation

One way to understand biochemical reactions is via graphical representation, which can provide more information and is easier to analyze than the reac-

tion list, especially for cases with a large number of reactions. For example, the global properties of molecular networks such as degree distribution and betweenness centrality can be more easily obtained from graphical representations than from the reaction list. Moreover, many theoretical results, *e.g.*, by graph theory and complex network theory, can be adopted to investigate these networks in order to obtain qualitative or quantitative predictions. In this book, we use three graphic representations to describe biochemical reactions or molecular interactions, depending on the models, *i.e.*, interaction graphs, incidence graphs, and species-reaction graphs. Detailed descriptions of the three representations are provided in Chapter 6.

2.3.1 Example of Interaction Graphs

First, consider the simple enzyme kinetic reaction (2.25) as an example of the interaction graph. The graphical representation is shown in Figure 2.2. Such a diagram represents the relationship or interactions between the components in a biomolecular system and is known as an interaction graph. As described in detail in Chapters 3 and 6, in an interaction graph, each node represents the concentration or the number of a chemical. Each edge or connection can be a linear or nonlinear function of the connected node. A positive value of the edge, *e.g.*, $A \to B$, implies that chemical A enhances the synthesis of chemical B, while a negative value implies that chemical A represses the synthesis of chemical B, which is also represented by $\dashv$ in contrast to $\to$.

Consider another example of a gene regulatory system with two genes in a eukaryotic cell. The biochemical reactions are shown below, where D_x and D_y denote the promoter regions of genes x and y, respectively. The multimerization (2.26), (2.30), (2.31); transportation (from cytoplasm to nucleus) (2.27), (2.32); and binding reactions of TFs to promoter regions (2.28), (2.29), (2.33) are described as follows:

$$P_x + P_x \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P_{2x}, \tag{2.26}$$

$$P_{2x} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} P'_{2x}, \tag{2.27}$$

$$P'_{2x} + D_y \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} P'_{2x}D_y, \tag{2.28}$$

$$P'_{2x} + P'_{2x}D_y \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} P'_{2x}\ P'_{2x}D_y, \tag{2.29}$$

$$P_y + P_y \underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} P_{2y}, \tag{2.30}$$

$$P_{2y} + P_{2y} \underset{k_{-6}}{\overset{k_6}{\rightleftharpoons}} P_{4y}, \tag{2.31}$$

$$P_{4y} \underset{k_{-7}}{\overset{k_7}{\rightleftharpoons}} P'_{4y}, \tag{2.32}$$

$$P'_{4y} + D_x \underset{k_{-8}}{\overset{k_8}{\rightleftharpoons}} P'_{4y}D_x, \tag{2.33}$$

where P_{2x}, P_{4y}, and P'_{2x}, P'_{4y} denote protein dimers and tetramers in the cytoplasm and the nucleus, respectively. $P'_{2x}D_y, P'_{4y}D_x$ and $P'_{2x}P'_{2x}D_y$ are all complexes. All these reactions are generally fast and occur within less than a few seconds.

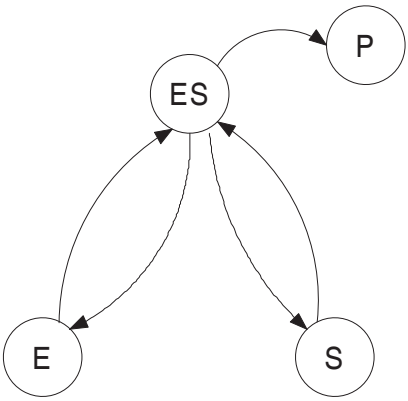

Figure 2.2 Graphical representation of the enzyme kinetic reaction via an interaction graph. All interactions or edges are positive

The reactions for the transcription of mRNAs (m_x, m_y), translation of proteins (P_x, P_y), and degradation of proteins and mRNAs are represented as follows:

$$D_x \overset{k_{mx0}}{\rightarrow} m_x + D_x, \tag{2.34}$$

$$P'_{4y}D_x \overset{k_{mx1}}{\rightarrow} m_x + P'_{4y}D_x, \tag{2.35}$$

$$m_x \overset{k_{Px}}{\rightarrow} P_x + m_x, \tag{2.36}$$

$$D_y \overset{k_{my0}}{\rightarrow} m_y + D_y, \tag{2.37}$$

$$P'_{2x}D_y \overset{k_{my1}}{\rightarrow} m_y + P'_{2x}D_y, \tag{2.38}$$

$$P'_{2x}\ P'_{2x}D_y \overset{k_{my2}}{\rightarrow} m_y + P'_{2x}\ P'_{2x}D_y, \tag{2.39}$$

$$m_y \xrightarrow{k_{Py}} P_y + m_y, \tag{2.40}$$

$$m_x \xrightarrow{d_{mx}} 0, \tag{2.41}$$

$$P_x \xrightarrow{d_{Px}} 0, \tag{2.42}$$

$$m_y \xrightarrow{d_{my}} 0, \tag{2.43}$$

$$P_y \xrightarrow{d_{Py}} 0. \tag{2.44}$$

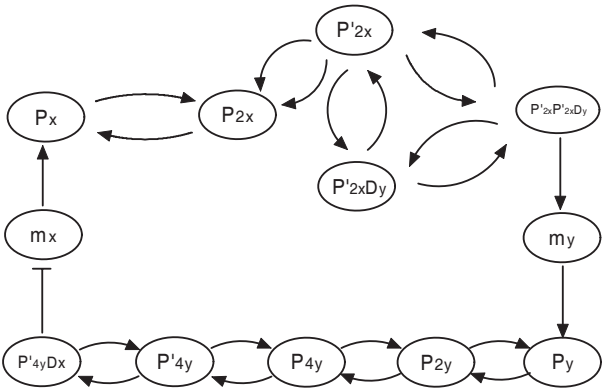

Figure 2.3 Schematic illustration of the graphical representation of a gene network by an interaction graph. '⊣' means a negative regulation, *i.e.*, repression, and '→' implies a positive regulation, *i.e.*, activation

Transcription and translation reactions are a chain of reactions, which are considerably slower than other reactions and generally require more than minutes to complete. The schematic representation for a gene network or gene regulatory network as an interaction graph is shown in Figure 2.3, where '⊣' means a negative regulation, *i.e.*, repression, and '→' implies a positive regulation, *i.e.*, activation. In this book, when there is no specific note on each edge of a network or a graph, a positive regulation (or interaction) and a negative regulation (or interaction) are represented by '→' and '⊣', respectively. However, the regulation between the two chemicals is positive or negative if the edge of the regulation is specifically indicated by '+' or '−'. The regulatory interactions between all the components are explicitly expressed in the associated diagram. When there are multiple feedback loops in a molecular network, it becomes difficult to gain some direct insight into the regulatory relations from the reactions. Even for this simple network, the interactions from p_{2y} to p_{2x} are not straightforward. However, the repression relation from p_{2y} to p_{2x} can easily be obtained from the graphical representation, as schematically illustrated in Figure 2.3.

2.3.2 Example of Incidence Graphs

To analyze biomolecular networks with control inputs, it is convenient to use incidence graphs, which are the same as the interaction graphs, except the two extended input and output nodes. Hence, compared with an interaction graph with n nodes, an incidence graph has $n+2$ nodes. All definitions in an incidence graph are the same as those of the corresponding interaction graph. When the input and output nodes are viewed equivalently as other nodes, it becomes an interaction graph. A detailed description of the incidence graph is provided in Chapter 6. For the example of (2.25), assuming the substrate S and the product P to be input and output of the system, respectively, the incidence graph can be simply represented by Figure 2.4, which has two additional nodes (input and output nodes) compared with the interaction graph of Figure 2.2. The incidence graph is used to study the relation between input and output.

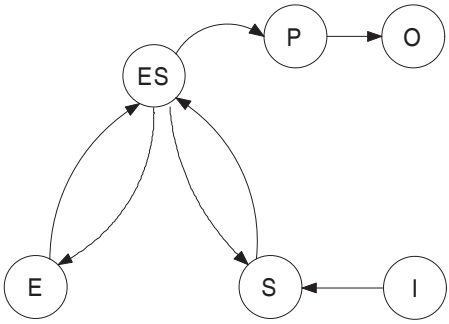

Figure 2.4 Graphical representation of the enzyme kinetic reaction by an incidence graph. All the definitions are the same as those of Figure 2.2, except input and output nodes. I and O represent the input and output nodes, respectively

2.3.3 Example of Species-reaction Graphs

In addition to interaction graphs and incidence graphs for the analysis of gene regulatory networks, SR graphs are used mainly in metabolic networks. An SR graph is composed of chemicals and reactions and is a general technique for studying the relationship between reaction network structures and the capacity for multistability. Unlike interaction graphs and incidence graphs which are directed graphs, an SR graph is an undirected graph. A detailed description of incidence graphs and SR graphs is provided in Chapter 6. For the example of the enzyme reaction (2.25), the SR graph is illustrated in Figure 2.5, where each circle represents a chemical (or a species) and each box stands for a reaction. Each edge or arc is labeled with the name of the

complex in which the species appears. An SR graph is mainly used to study the multistability in a molecular network.

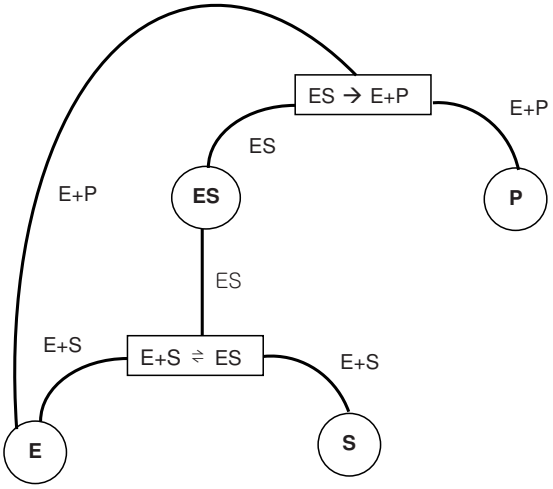

Figure 2.5 Graphical representation of an enzyme kinetic reaction by an SR graph

2.4 Biochemical Kinetics

Molecular networks can be modeled and analyzed as stochastic or deterministic systems. Different approaches may highlight different aspects of the same list of reactions. Moreover, even for a given model, many different aspects can be revealed from theoretical and numerical analysis of the network.

In dynamical modeling, the reaction rates that depend on the change of participating species over time need to be defined, *e.g.*, in terms of the number or the concentration of each molecule. In order to obtain the temporal behavior of the species, the functional relation of change in the number or the concentration need to be specified, depending on the approach chosen. Specifically, for each reaction, we identify its rate constant, which depends on temperature and specifies the amount of time that the reaction takes, and an associated rate law, which specifies the amount of state changes or the probability that the reaction occurs in a small time interval.

For example, consider a simple reaction of the form

$$A \xrightarrow{k} B. \tag{2.45}$$

This reaction means that A is transformed to B at a rate of k. In the deterministic formulation, the reaction specifies the state changes in which the

concentration of A decreases concomitantly with the increase in the concentration of B. The amount of state change in a small time interval dt is given by $k[A]dt$, where $[A]$ is the concentration of A. On the other hand, in the stochastic formulation, the reaction specifies the state changes in which the total number of A decreases by one, and the total number of B increases by one. The probability for this reaction is given by $k\{\#A\}dt$, where $\{\#A\}$ is the number of A present. Note that in this book, $[A]$ represents the concentration of species A, but sometimes A also stands for the concentration of species A to simplify the representation. For the same reason, we also occasionally drop # to represent the number of species A directly by using A instead.

When modeling biochemical kinetics, both the stochastic and deterministic approaches follow the mass action law, which states that the reaction rate is proportional to the probability of a collision of the reactants. This probability is in turn proportional to the concentrations of reactants to the power of the molecularity, *i.e.*, the number in which they enter the specific reaction. There are some minor differences between stochastic and deterministic rate constants. In order to conduct a stochastic simulation, the deterministic rate constants must be converted in an appropriate way to stochastic rate constants.

Next, we will provide a general mathematical framework of both stochastic and deterministic approaches for modeling molecular networks. Especially, we will show how the master equations, Langevin (stochastic differential) equations, cumulant equations, and then deterministic differential equations can be obtained to model biochemical kinetics and further provide a brief comparison among them.

2.5 Stochastic Representation

A cellular system is well regulated at a molecular level but is an inherently noisy process, from transcriptional control, alternative splicing, translation, diffusion to chemical modification reactions, all of which involve stochastic fluctuations. Such stochastic noise may not only affect the dynamics of biological systems but may also be exploited by living organisms to actively facilitate certain functions. From an evolutionary viewpoint, noise is also assumed to be used for cellular and population variability control. Due to very low copy numbers for many species in living cells, the origins of stochasticity can be traced to the random transitions among discrete chemical states, which implies that a model of a molecular network should be able to present this discrete nature of small numbers both qualitatively and quantitatively. To capture the discrete and stochastic nature of biochemical kinetics with low concentrations or small numbers of molecules, a stochastic modeling formulation can be adopted to describe such a biological system. In stochastic modeling, one can estimate the time of each reaction from statistical properties. For example, stochastic simulation provides the time of a reaction from

a probability distribution. The result of a simulation run is one possible realization of the temporal evolution of the system.

The stochastic framework considers the exact numbers of molecules present, which are discrete quantities. Such a strategy may identify how many molecules of each component are present in the system. The stochastic framework grasps the essence of the stochastic collision of biochemical components, *i.e.*, the components change discretely, but which change occurs and when it occurs are probabilistic. Actually, a cellular system at a molecular level can be considered to be governed by stochastic processes with many random events, or by a continuous time and discrete-space Markov chain.

Consider, for example, a DNA binding protein P, whose binding to and unbinding from DNA follow elementary reactions

$$P + DNA \overset{k_+}{\rightharpoonup} P \cdot DNA, \tag{2.46}$$

$$P \cdot DNA \overset{k_-}{\rightharpoonup} P + DNA. \tag{2.47}$$

This list of reactions involves chemicals of three types: P, DNA, and $P \cdot DNA$, which is a complex of P and DNA. There are multiple possible reactions involved in (2.46)–(2.47), and the states change discretely when any one of the reactions occurs. Assume that the system state at time t is $(\{\#P\}, \{\#DNA\}, \{\#P \cdot DNA\})$. After a short time d$t$, the state will change to $(\{\#P - 1\}, \{\#DNA - 1\}, \{\#P \cdot DNA + 1\})$ with the probability $k_+\{\#P\}\{\#DNA\}\mathrm{d}t$, to state $(\{\#P+1\}, \{\#DNA+1\}, \{\#P \cdot DNA-1\})$ with the probability $k_-\{\#P \cdot DNA\}\mathrm{d}t$, or stay in the same state with the probability $1 - k_+\{\#P\}\{\#DNA\}\mathrm{d}t - k_-\{\#P \cdot DNA\}\mathrm{d}t$, where $\{\#A\}$ indicates the number of molecule A (Gibson and Mjolsness 2001).

In stochastic modeling, one can also deal with the probabilities rather than the numbers of molecules. In other words, the state or variable is the probability distribution over all configurations. A configuration is a list of numbers of molecules. The dynamics of probabilities obey a master equation, a linear differential equation, which describes the temporal evolution of the probability distribution (Van Kampen 1992) for the discrete transitions of molecules.

2.5.1 Master Equations for a General Molecular Network

The master equations description accounts for the probabilistic nature of cellular biochemical processes and can be viewed as a continuous time discrete-space Markov model. It describes the time evolution of the probability of having a certain number of molecules, and its result is usually taken as a gold standard for numerical simulation in computational biology due to its detailed representation and also due to the lack of experimental data. In the master equation, reaction rates are transformed into probability transition rates.

Suppose that the number of molecules of reactant X can be any positive integer $\{\#X\} = n$. Let $P_n(t)$ denote the probability that there are

n molecules of X at time t. Then, we want to know the temporal evolution of the probability which governs its development. In other words, we want to express $P_n(t + \mathrm{d}t)$ in terms of $P_m(t)$ for all m, *i.e.*, to express the probability of having n molecules at time $t + \mathrm{d}t$ in terms of the probability of all possible values for m molecules at time t. The occurrence of event $\{\#X(t+\mathrm{d}t)\} = n$ can be thought of as the the occurrence of mutually exclusive event $(\{\#X(t+\mathrm{d}t)\} = n, \{\#X(t)\} = m)$ for all possible m. When taking the conditional probabilities into account, we have

$$\begin{aligned} P(\{\#X(t+\mathrm{d}t)\} = n) &= \sum_m P(\{\#X(t+\mathrm{d}t)\} = n, \{\#X(t)\} = m) \\ &= \sum_m P_{m,n} P(\{\#X(t)\} = m), \end{aligned} \tag{2.48}$$

where $P_{m,n} = P(\{\#X(t + \mathrm{d}t)\} = n|\{\#X(t)\} = m)$ is the transition probability of changing from m to n molecules in the time interval $\mathrm{d}t$ for $m, n = 0, 1, 2, ...$ The summation of (2.48) is over all possible m.

The stochastic representation (2.48) obeys a Markov process because the transition probability itself neither explicitly depends upon the time at which the transition occurs, nor does it depends on the path on which the change occurred. Taking the limit as $\mathrm{d}t \to 0$ in the difference $P(\{\#X(t + \mathrm{d}t)\} = n) - P(\{\#X(t)\} = n)$ leads to a differential equation in the probabilities, *i.e.*, the master equation.

An Illustrative Example of the Master Equation

In order to clearly explain the master equation, consider the reactions (2.46)–(2.47). Let $X = (X_1, X_2, X_3) = (\{\#P\}, \{\#DNA\}, \{\#P \cdot DNA\})$ be the numbers of the three species. Define $P(X; t)$ to be the probability function for the state $(\{\#P\}, \{\#DNA\}, \{\#P \cdot DNA\}) = X$ at time t. The probability of being in state $X = (X_1, X_2, X_3)$ at time $t + \mathrm{d}t$ is composed of the sum of terms which describe all possible previous states multiplied by their respective transition probabilities. Then, the probability at $t + \mathrm{d}t$ is given by

$$\begin{aligned} P(X; t+\mathrm{d}t) = {} & P_{(X_1+1,X_2+1,X_3-1),(X_1,X_2,X_3)} P(X_1+1, X_2+1, X_3-1; t) \\ & + P_{(X_1-1,X_2-1,X_3+1),(X_1,X_2,X_3)} P(X_1-1, X_2-1, X_3+1; t) \\ & - P_{(X_1,X_2,X_3),(X_1-1,X_2-1,X_3+1)} P(X_1, X_2, X_3; t) \\ & - P_{(X_1,X_2,X_3),(X_1+1,X_2+1,X_3-1)} P(X_1, X_2, X_3; t) \\ & + P(X_1, X_2, X_3; t). \end{aligned} \tag{2.49}$$

The transition probability is assumed to be proportional to their numbers and time interval dt

$$P_{(X_1+1,X_2+1,X_3-1),(X_1,X_2,X_3)} = k_+(X_1+1)(X_2+1)\mathrm{d}t, \tag{2.50}$$
$$P_{(X_1-1,X_2-1,X_3+1),(X_1,X_2,X_3)} = k_-(X_3+1)\mathrm{d}t, \tag{2.51}$$
$$P_{(X_1,X_2,X_3),(X_1-1,X_2-1,X_3+1)} = k_+X_1X_2\mathrm{d}t, \tag{2.52}$$
$$P_{(X_1,X_2,X_3),(X_1+1,X_2+1,X_3-1)} = k_-X_3\mathrm{d}t. \tag{2.53}$$

Substituting (2.50)–(2.53) in (2.49), taking $P(X_1, X_2, X_3; t)$ to the left-hand side, and dividing by $\mathrm{d}t$, and then considering the limit as $\mathrm{d}t \to 0$, we obtain a differential equation in the probability as follows:

$$\begin{aligned}\frac{\mathrm{d}P(X;t)}{\mathrm{d}t} &= \lim_{dt\to 0}\frac{P(X_1,X_2,X_3;t+\mathrm{d}t)-P(X_1,X_2,X_3;t)}{\mathrm{d}t}\\ &= k_+(X_1+1)(X_2+1)P(X_1+1,X_2+1,X_3-1;t)\\ &\quad +k_-(X_3+1)P(X_1-1,X_2-1,X_3+1;t)\\ &\quad -(k_+X_1X_2+k_-X_3)P(X_1,X_2,X_3;t).\end{aligned} \tag{2.54}$$

An equation of the type (2.54) is called the master equation, which is a continuous-time discrete-space Markov chain because the probability of the next transition depends on the current state only, and not on the history of states. The first two terms are the gain of state $X = (X_1, X_2, X_3)$ due to transitions from other states, and the last term (*i.e.*, $(k_+X_1X_2 + k_-X_3)P(X_1, X_2, X_3; t)$) is the loss due to transitions from $X = (X_1, X_2, X_3)$ to other states.

A General Molecular Network

Consider a general molecular network with N molecular species $\{S_1, ..., S_N\}$ that react through M reaction channels $\{R_1, ..., R_M\}$. Let $X = (X_1, ..., X_N)$ be the state of the molecules at time t, *i.e.*, X_i is the number of the ith molecule at time t. Define $P(X; t)$ to be the probability function for the state X at time t. Here, we occasionally drop t to simplify the expression in the following, *e.g.*, $X = X(t)$, without confusion. Then, the temporal evolution of the system is governed by a set of reactions $X \to X'$ and $X' \to X$. The probability $P(X; t)$ that the system is in state X at time t evolves according to the master equation (Van Kampen 1992)

$$\frac{\partial P(X;t)}{\partial t} = \sum_{X'}\Big\{P_{X'X}P(X';t) - P_{XX'}P(X;t)\Big\}, \tag{2.55}$$

where $P_{X'X}$ is the transition probability from state X' to state X, and $P_{XX'}$ is the transition probability from state X to state X'. The master equation is nothing but a gain–loss equation for the probability distribution $P(X; t)$ or a stochastic birth-and-death process. The summation of (2.55) is for all possible states of X'. The first term is the gain of state X due to transitions from other states X' to X, and the second term is the loss due to transitions from X to other states.

Define $a_j(X)\mathrm{d}t$ to be the probability, given X, that one R_j reaction occurs in the next infinitesimal time interval $[t, t + \mathrm{d}t)$ and ν_{ji} to be the change in the number of S_i molecules produced by one R_j reaction ($j = 1, ..., M$ and $i = 1, ..., N$). The master equation (2.55) for a general molecular network can be rewritten as

$$\frac{\partial P(X;t)}{\partial t} = \sum_{j=1}^{M} \Big\{ a_j(X - \nu_j)P(X - \nu_j; t) - a_j(X)P(X;t) \Big\}, \tag{2.56}$$

where $\nu_j = (\nu_{j1}, ..., \nu_{jN})$. Note that the state X is discrete but the probability distribution P is continuous.

Generally, the master equation itself is not analytically or numerically solvable in any but the simplest cases. Therefore, one has to resort to Monte Carlo types of simulations that produce a random walk through the possible states of the system. In other words, instead of calculating the probability distribution $P(X;t)$, the approaches simulate the time evolution of a particular trajectory, starting at an initial state X_0. Various methods have been developed, such as the Gillespie stochastic simulation algorithm (SSA) (Gillespie 1976, Gillespie 1977). Recently, some more computationally efficient algorithms have been developed such as the Gibson–Bruck algorithm (Gibson and Bruck 2000), the approximate simulation strategies (Gillespie 2001, Gillespie and Petzold 2003), the hybrid simulation strategies (Kiehl *et al.* 2004, Rao and Arkin 2003, Puchalka and Kierzek 2004), and parallel simulation strategies (Chen *et al.* 2005). However, making the stochastic simulation algorithms more efficient and applying stochastic simulation algorithms to larger models are still the subject of active research. For example, the perfect sampling algorithm was proposed to recast Gillespie's SSA in the light of Markov chain Monte Carlo methods and combine it with the dominated coupling from past algorithms in order to provide guaranteed sampling from the stationary distribution (Hemberg and Barahona 2007).

Equilibrium Probability Distribution

In the stochastic framework, it is difficult to define an equilibrium configuration in terms of the numbers of molecules as in a deterministic framework since any reaction which changes the numbers of molecules takes places in probability. In deterministic differential equations, an equilibrium means that there is no net change in the numbers or concentrations of molecules. The concept of the equilibrium suffices for differential equation models, but not for stochastic models. In contrast, for stochastic models, there is equilibrium probability distribution, *i.e.*, a set of probabilities that the system has certain numbers of molecules, even though the numbers of molecules may change.

Generally, it is difficult to get equilibrium probability distribution of a master equation analytically, if not impossible, although some statistical quantities such as means and variances can be obtained based on it. However, for some

simple cases, it can be obtained by recursion relations. For example, consider a set of coupled reactions called the 'Lotka–Volterra' system,

$$X \xrightarrow{k_1} 2X, \tag{2.57}$$

$$X + Y \xrightarrow{\beta} 2Y, \tag{2.58}$$

$$Y \xrightarrow{k_2} 0. \tag{2.59}$$

The time evolution of the joint probability $P(X, Y; t)$ obeys the following master equation:

$$\begin{aligned}\frac{\partial P(X,Y;t)}{\partial t} &= k_1(X-1)P(X-1,Y;t) - k_1 X P(X,Y;t) \\ &\quad + k_2(Y+1)P(X,Y+1;t) - k_2 Y P(X,Y;t) \\ &\quad + \beta(X+1)(Y-1)P(X+1,Y-1;t) - \beta XY P(X,Y;t).\end{aligned} \tag{2.60}$$

To obtain its equilibrium distribution, we define

$$A(X;t) = \sum_{Y=0}^{\infty} P(X,Y;t) \quad \text{and} \quad B(Y;t) = \sum_{X=0}^{\infty} P(X,Y;t), \tag{2.61}$$

where $A(X;t)$ is the marginal probability of X at time t, irrespective of the number of Y. The probability $B(Y;t)$ is defined analogously.

Supposing that there is equilibrium probability distribution, one can obtain

$$\dot{A}(X;t) = 0 \quad \text{for} \quad Y = 0, 1, 2, ..., \tag{2.62}$$

$$\dot{B}(Y;t) = 0 \quad \text{for} \quad X = 0, 1, 2, ..., \tag{2.63}$$

where $\dot{A} = \mathrm{d}A/\mathrm{d}t$ and $\dot{B} = \mathrm{d}B/\mathrm{d}t$. Then, using recursion relations, we can obtain the equilibrium distribution

$$P(X,Y;t) = 0 \ \text{ for all } \ X = 0, 1, ... \ \text{ and } \ Y = 1, 2, ... \tag{2.64}$$

To derive (2.64) in more detail, on the basis of the definition of $B(Y;t)$ in (2.61), by summing (2.60) over all X, we have

$$\begin{aligned}\dot{B}(Y;t) &= k_2(Y+1)B(Y+1;t) - k_2 Y B(Y;t) - \beta Y \sum_{X=0}^{\infty} X P(X,Y;t) \\ &\quad + \beta(Y-1) \sum_{X=0}^{\infty} X P(X,Y-1;t).\end{aligned} \tag{2.65}$$

Then, we have

$$\begin{aligned} 0 = \dot{B}(0,t) &= k_2(0+1)B(0+1,t) - k_2 \cdot 0 \cdot B(0,t) \\ &\quad +\beta(0-1)\sum_{X=0}^{\infty} XP(X,0-1,t) \\ &\quad -\beta \cdot 0 \cdot \sum_{X=0}^{\infty} XP(X,0,t) \\ &= k_2 B(1,t). \end{aligned} \tag{2.66}$$

Note that $P(X,-1,t) = 0$ because all X, Y are non-negative numbers. Hence, we can obtain

$$B(1,t) = \sum_{X=0}^{\infty} P(X,1,t) = 0. \tag{2.67}$$

Because all the probabilities are non-negative, we have

$$P(X,1,t) = 0 \quad \text{for} \quad X = 1,2,.... \tag{2.68}$$

Proceeding with this process, we can obtain (2.64). Similarly, by using

$$\begin{aligned} \dot{A}(X;t) &= k_1(X-1)A(X-1;t) - k_1 XA(X;t) - \beta X \sum_{Y=0}^{\infty} YP(X,Y;t) \\ &\quad +\beta(X+1)\sum_{Y=0}^{\infty} YP(X+1,Y;t), \end{aligned} \tag{2.69}$$

we obtain

$$P(X,Y;t) = 0 \quad \text{for all} \quad X = 1,2,... \quad \text{and} \quad Y = 0,1,.... \tag{2.70}$$

Integrating (2.64) and (2.70), it can be obtained that the only non-zero probability left in the equilibrium distribution is

$$P(0,0;t) = 1. \tag{2.71}$$

Therefore, the stochastic 'Lotka–Volterra' system does not have equilibrium distribution other than the one in which both species X and Y are extinct (Reddy 1975), *i.e.*, $X = Y = 0$. However, on the basis of (2.57)–(2.59), the deterministic 'Lotka–Volterra' system can be written as

$$\frac{\mathrm{d}X}{\mathrm{d}t} = k_1 X - \beta XY, \tag{2.72}$$

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = \beta XY - k_2 Y. \tag{2.73}$$

In this book, we occasionally use $\dot{X}$ to represent $\mathrm{d}X/\mathrm{d}t$. Clearly, it has an equilibrium $\bar{X} = k_2/\beta$ and $\bar{Y} = k_1/\beta$, which are generally non-zero. The difference arises from the different possibilities for the identification of birth and death terms in the deterministic equation. Another example where the equilibrium probability distribution can be solved analytically is the model for the signaling cycle (Levine *et al.* 2007).

2.5.2 Stochastic Simulation

Although the analytical solution of the master equation is rarely available, the density function can be constructed numerically using the SSA (Gillespie 1976). Generally, the SSA first constructs numerical realizations of $X(t)$ and then averages the results of many realizations. Specifically, we compute the reaction probability density function $P(\Delta t, \mu|X;t)$, which is the joint probability density function of two random variables, the time to the next reaction Δt and the index of the next reaction μ, given X. The reaction probability density function for the general molecular network (2.56) takes the form

$$P(\Delta t, \mu|X;t) = a_\mu(X)\exp(-a_0(X)\Delta t) \quad (\Delta t \geq 0;\ \mu = 1, ..., M), \tag{2.74}$$

with

$$a_0(X) = \sum_{j=1}^{M} a_j(X). \tag{2.75}$$

The reaction probability density function provides the basis for the SSA. According to the joint density function (2.74), the next reaction and the time of its occurrence can be generated through the direct method as follows. Draw two random numbers r_1 and r_2 from a uniform distribution in unit interval $[0, 1)$. The time to the next reaction Δt and the index of the next reaction μ, given X, can be taken as follows:

$$\Delta t = \frac{1}{a_0(X)}\ln(\frac{1}{r_1}), \tag{2.76}$$

$$\mu = \text{the smallest integer satisfying } \sum_{j'=1}^{\mu} a_{j'}(X) > r_2 a_0(X). \tag{2.77}$$

The Gillespie direct method for exact simulation of the master equation (2.56) is shown in Table 2.1. Despite its simplicity in expression, the computational cost of the Gillespie algorithm increases drastically with the number of reactions and the number of molecules. The increment in computational cost is primarily due to the generation of the two random numbers and the enumeration of reactions to determine the next reaction. For example, when the number of molecules is equal to 10^6, the time step Δt will become excessively small, *i.e.*, on the order of 10^{-6}, which means that more time steps or iterations are needed.

For illustrative purposes, we use the example for the Gillespie algorithm presented in (Gillespie 2001). There are three molecular species (S_1, S_2, S_3) and four reaction channels as follows:

$$S_1 \xrightarrow{c_1} 0, \tag{2.78}$$

$$S_1 + S_1 \xrightarrow{c_2} S_2, \tag{2.79}$$

$$S_2 \xrightarrow{c_3} S_1 + S_1, \tag{2.80}$$

$$S_2 \xrightarrow{c_4} S_3. \tag{2.81}$$

Table 2.1 Gillespie direct method for exact simulation of (2.56) (Gillespie 1977)

Step 1. Initialization: set $t = 0$ and fix the initial numbers of molecules X_0.
Step 2. Calculate the propensity function a_j, $j = 1, ..., M$.
Step 3. Generate two random numbers r_1 and r_2 in $[0, 1)$.
Step 4. Determine Δt and μ according to (2.76)–(2.77).
Step 5. Execute reaction μ and advance time Δt, *i.e.*, $t \leftarrow t + \Delta t$. If t reaches T_{max}, terminate the computation. Otherwise, go to Step 2.

For these reactions, the reaction propensities, which describe the probability of one molecule colliding with another, take the forms

$$a_1 = c_1 X_1, \tag{2.82}$$
$$a_2 = c_2 X_1 (X_1 - 1)/2, \tag{2.83}$$
$$a_3 = c_3 X_2, \tag{2.84}$$
$$a_4 = c_4 X_2, \tag{2.85}$$

where X_1 and X_2 are the numbers of molecules S_1 and S_2, respectively.

The simulation results are shown in Figure 2.6 when using the rate constant values

$$c_1 = 1, \quad c_2 = 0.002, \quad c_3 = 0.5, \quad c_4 = 0.04 \tag{2.86}$$

and the initial conditions

$$X_1 = 10^5, \quad X_2 = X_3 = 0, \tag{2.87}$$

where X_3 is the number of molecules of S_3.

Time delay exists in many cellular processes and in a mathematical model, the effect of the delay may be drastic. When delay times are significant, both analytical and numerical modeling should take into account their crucial influences. To explore the combined effects of time delay and intrinsic noise on cellular dynamics, a modified Gillespie method was developed (Bratsun *et al.* 2005), which allows incorporation of delayed reactions and accounts for the non-Markovian properties of random biochemical events with time delay. The formal steps for the modified method are shown in Table 2.2.

To account for the modified method more clearly, we use a simple model presented in (Bratsun *et al.* 2005) as an example. Suppose that the protein can exist both in the form of monomers X_1 and dimers X_2, and protein production occurs with a time lag τ and can only occur if the operator D is unoccupied at time t. The reaction channels are as follows:

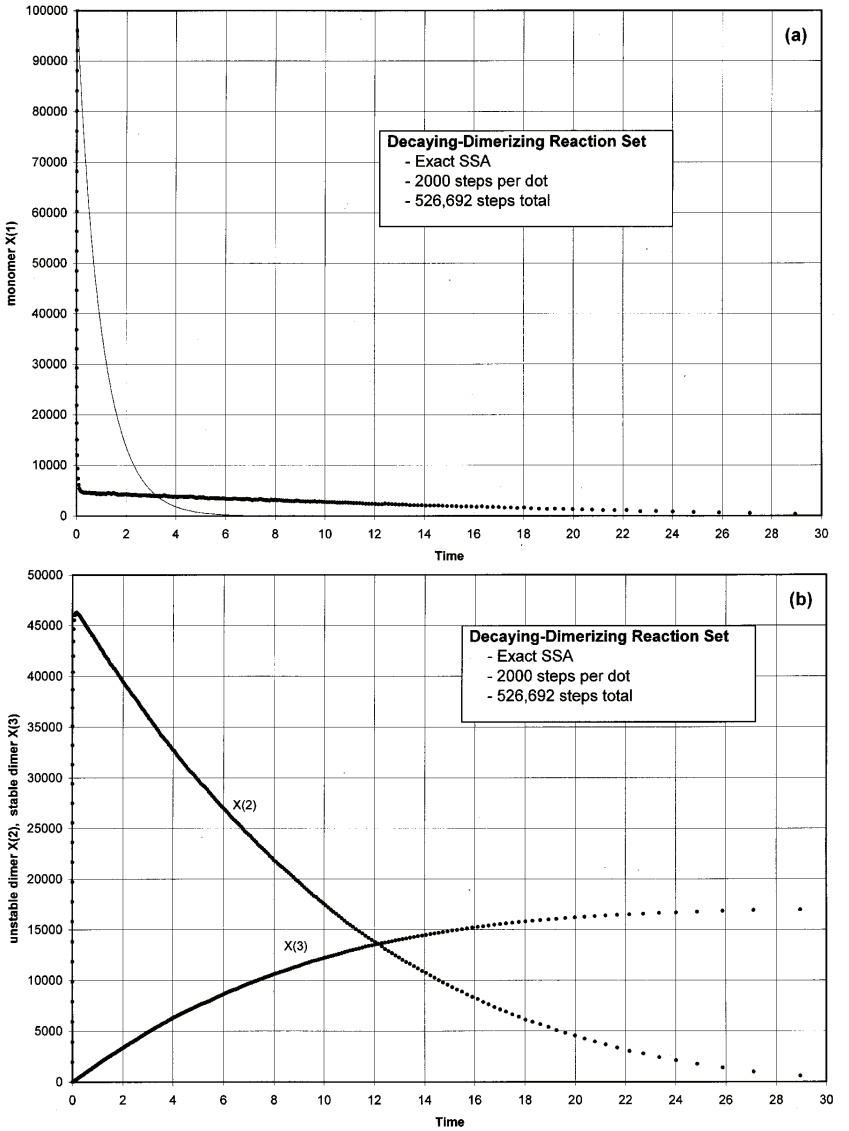

Figure 2.6 Stochastic simulation of the system (2.78)–(2.81) for the rate constants (2.86) and the initial condition (2.87) (from (Gillespie 2001))

$$D_0 \xrightarrow{k_0} D_0 + X_1, \tag{2.88}$$

$$X_1 + X_1 \xrightarrow{k_2} X_2, \tag{2.89}$$

$$X_2 \xrightarrow{k_{-2}} X_1 + X_1, \tag{2.90}$$

$$D_0 + X_2 \xrightarrow{k_1} D_1, \tag{2.91}$$

$$D_1 \xrightarrow{k_{-1}} D_0 + X_2, \tag{2.92}$$

$$X_1 \xrightarrow{d} 0, \tag{2.93}$$

Table 2.2 Modified Gillespie method to incorporate delayed reactions (Bratsun *et al.* 2005)

Step 1. Initialization: set $t = 0$ and fix the initial numbers of molecules X_0 and the reaction counter.
Step 2. Calculate the propensity function a_j, $j = 1, ..., M$.
Step 3. Generate two random numbers r_1 and r_2 in $[0, 1)$.
Step 4. Compute the time interval until the next reaction according to (2.76).
Step 5. Check whether there are delayed reactions scheduled within time interval $[t, t + \Delta t]$. If yes, time t advances to the time $t_d = t + \tau$ of the next scheduled delayed reaction, states X are updated according to the delayed reaction channel, and the counter is increased $i \leftarrow i + 1$. Proceed to Step 2. Otherwise, go to Step 6.
Step 6. Find the channel of the next reaction μ according to (2.77).
Step 7. If the selected reaction μ is not delayed, update X by executing reaction μ, update time $t \leftarrow t + \Delta t$, and increase counter $i \leftarrow i + 1$. If the selected reaction is delayed, update is postponed until time t_d. If t reaches T_{max}, terminate the computation. Otherwise, go to Step 2.

where D_0 and D_1 are the unoccupied operator and the occupied operator by repressor X_2, respectively.

When utilizing the inherent separation of time scales and eliminating the fast variables X_2 and D_1 corresponding to reactions (2.89)–(2.92), the deterministic equation for the number of monomers X_1 takes the form

$$\left(1 + 4\epsilon X_1 + \frac{4\epsilon\delta X_1}{(1 + \epsilon\delta X_1^2)^2}\right)\frac{\mathrm{d}X_1}{\mathrm{d}t} = \frac{k_0}{1 + \epsilon\delta X_1^2(t - \tau)} - \mathrm{d}X_1, \qquad (2.94)$$

where $\epsilon = k_1/k_{-1}$ and $\delta = k_2/k_{-2}$. (2.94) is an approximation of the original system using the quasi-steady-state equilibrium of (2.89)–(2.92), but clearly, deterministic simulation of (2.94) is considerably simpler. The stochastic simulation results are shown in Figure 2.7. See (Bratsun *et al.* 2005) for more details on the modified Gillespie method.

Although the equilibrium probability distribution of the master equation (2.55) is difficult to derive analytically, it can be calculated numerically on the basis of the following algebraic equation:

$$\sum_{X'}\Big\{P_{X'X}P(X';t) - P_{XX'}P(X;t)\Big\} = 0. \qquad (2.95)$$

For instance, by enumerating all possible (feasible) configurations or states of X and then substituting them into (2.95), we have simultaneous equations

for all feasible states. Then, exploring the sparse structure of these linear equations, *e.g.*, using the Arnoldi method (Cao and Liang 2008), we can estimate the equilibrium probability distribution of P for all states or configurations of X.

The exact stochastic simulation algorithms are computationally expensive, and thereby only realistic for the computation of a small-scale system. To simulate a large-scale biomolecular system, approximation schemes are usually adopted, *e.g.*, the finite state projection approach, the parallel computation scheme, the adaptive time-step procedure, and the model reduction (or aggregation) method based on time-scale separation or continuous–discrete variable separation, by exploiting the special dynamical or topological properties of the concerned system.

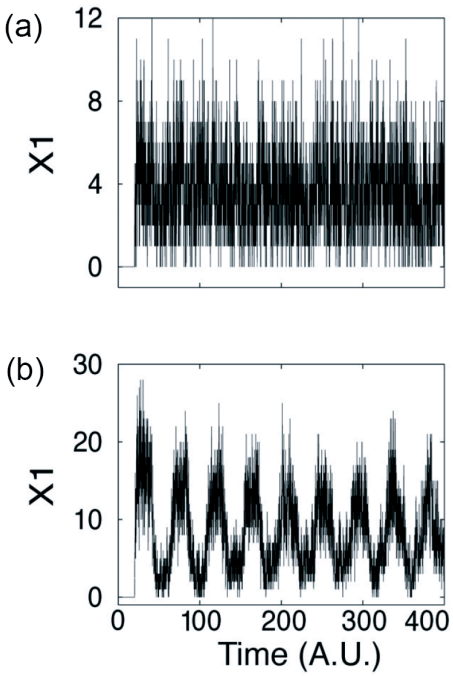

Figure 2.7 Stochastic simulation of the system (2.88)–(2.93). The parameter values are $\epsilon = 0.1$, $\delta = 0.2$, $k_1 = 100$, $k_{-1} = 1000$, $k_2 = 200$, $k_{-2} = 1000$, $d = 4$, $\tau = 20$, $k_0 = 20$ (a), and $k_0 = 70$ (b) (from (Bratsun *et al.* 2005))

2.5.3 Analysis of Sensitivity and Robustness of Master Equations

Sensitivity characterizes the ability of living organisms to adequately react to certain stimulus. It quantifies the dependence of system behavior on parameters. In deterministic modeling, robustness is usually quantified by calculating sensitivity, *e.g.*, period and amplitude sensitivity in quantifying robustness of circadian rhythms. Sensitivity analysis of stochastic systems has recently become popular due to its relevance to the simulation of cellular processes.

Using an analogue of the classical sensitivity analysis, the parameter sensitivity can also be applied to discrete stochastic dynamical systems. In stochastic systems, the state is probability distribution and parameters affect it indirectly through a master equation. The parameter sensitivity applied to the probability distribution is given by (Gunawan *et al.* 2005)

$$S_j(X;t) = \frac{\partial P(X(p);t)}{\partial p_j}, \tag{2.96}$$

where p denotes the parameter vector. In traditional sensitivity analysis, one implicit assumption is that the outputs are continuous with respect to parameters. However, in stochastic systems, the output takes random values with probability. A sensitivity measure for discrete stochastic systems is defined by (Gunawan *et al.* 2005)

$$s_j(t) = E\left[\left|\frac{\partial P(X;t)}{\partial p_j}\right|\right] = \int_X \left|\frac{\partial P(X;t)}{\partial p_j}\right| P(X;t)\mathrm{d}X, \tag{2.97}$$

where $E[\cdot]$ means the expected values on X. The absolute value is used here because only the magnitude of the sensitivity coefficient is concerned. Unlike the sensitivity coefficient of a single system state with respect to a parameter, the dependence of the system state X on parameter p is implicitly assumed.

The evolution of sensitivity coefficients can be derived from (2.55) by taking the derivatives with respect to parameters

$$\begin{aligned}\frac{\partial S_j(X;t)}{\partial t} = \sum_{X'}\Big\{ & P_{X'X}S_j(X';t) - P_{XX'}S_j(X;t) \\ & + \frac{\partial P_{X'X}}{\partial p_j}P(X';t) - \frac{\partial P_{XX'}}{\partial p_j}P(X;\,t)\Big\}.\end{aligned} \tag{2.98}$$

The evolution of sensitivity coefficients (2.98) should be solved simultaneously with its respective master equation (2.55). Generally, its analytical solution is difficult to construct. The SSA cannot be directly applied to solve the sensitivity equations. Techniques for estimating the sensitivity coefficients have been developed on the basis of a black-box approach such as finite difference. Specifically, the probability density function is constructed through a cumulative distribution function from the SSA. The cumulative distribution function is given by

$$F(X;t) = \int_{-\infty}^{X} P(\tilde{X};t)\mathrm{d}\tilde{X}, \tag{2.99}$$

from which the density function $P(X;t)$ can be obtained as follows:

$$P(X;t) = \frac{\mathrm{d}F(X;t)}{\mathrm{d}X}. \tag{2.100}$$

Then, the sensitivity coefficients can be numerically computed according to difference approximation as follows:

$$\begin{aligned} & S_j(X;t) = \frac{\partial P(X;t)}{\partial p_j} \\ = & \frac{P(X(p_j + \Delta p_j);t) - P(X(p_j - \Delta p_j);t)}{2\Delta p_j} \\ = & \frac{F(P(X(p_j + \Delta p_j));t) - 2F(P(X(p_j));t) + F(P(X(p_j - \Delta p_j));t)}{(\Delta p_j)^2} \end{aligned} \tag{2.101}$$

where Δp_j is the perturbation to p_j and should be small enough to minimize truncation error but large enough to accurately predict the effect of the parameter changes over the level of internal noise (Gunawan *et al.* 2005). Generally, this balance in choosing step sizes in the parameters may hinder the application of finite difference to the probability density function. Recently, spectral methods for parameter sensitivity in stochastic dynamical systems were developed (Kim *et al.* 2007). The authors used spectral polynomial chaos expansions to represent statistics of the system dynamics as polynomial functions of system parameters. Such an approach allows an accurate and robust evaluation of sensitivities, even for the case of large-magnitude parametric perturbations. In addition, it enables studies of the predictability of the system given uncertainty, variability, or external noise in the model parameters, and estimation of corresponding uncertainty of the predicted output state statistics.

2.5.4 Langevin Equations

The Langevin equational description to represent a cellular system explicitly incorporates the effects of noise. It takes the form of continuous differential equations augmented with additive or multiplicative stochastic terms, called stochastic differential equations. Due to the explicit incorporation of noise, the Langevin approach is ideal to describe constructive effects of stochastic fluctuations in cellular systems. Two typical examples of Langevin equations with additive and multiplicative stochastic terms take the forms

$$\dot{X} = f(X) + \Gamma(t) \tag{2.102}$$

and

$$\dot{X} = h(X) + g(X)\Gamma(t), \tag{2.103}$$

where $\Gamma(t)$ can be a Gaussian white noise or other fluctuating random term (Hasty *et al.* 2000). In (2.102), the noise term does not depend on the current state, whereas in (2.103), the noise term depends on the current state. Here, $g(X)$ is a deterministic function of X, which determines the scaling of the noise. Note that unlike a master equation, the state X in the Langevin equations takes continuous values, and hence the Langevin equations can be viewed as an approximation to the corresponding master equation, *i.e.*, they approximate discrete state variable X of the master equation by continuous variable X of (2.103).

In order to describe the fluctuations, one can generally proceed by the following steps (Van Kampen 1992) when $g(X)$ and Γ are not explicitly given.

1. Write the deterministic macroscopic equations of the system.

2. Add the Langevin force, *i.e.*, the second term of (2.102) or (2.103).

3. Adjust the specified constants related to the Langevin force so that the stationary solution reproduces the correct mean square fluctuations as known from statistical mechanics or other considerations.

The master equation and various stochastic simulation algorithms account for the stochastic and discrete nature of biochemical reactions and have been widely used to investigate the properties and effects of the internal noise as the gold standard of numerical computation. However, the computational efficiency rapidly degrades as the complexity of a system increases. In addition, a master equation cannot provide clear perspectives on the origins and magnitude of the internal noise. Staring from a master equation, when a system possesses a macroscopically infinitesimal time scale in the sense that during any time increment $\mathrm{d}t$ on that scale, all the reaction channels fire much more than once and none of the propensity functions change appreciably, its dynamics can be well approximated by Langevin equations (Gillespie 2000). In particular, when the number of each species is large (*e.g.*, over hundreds in a cell), Langevin equations can well describe the dynamics of cellular systems.

The Langevin equations of a general molecular network can be straightforwardly derived from (2.56) and take the form

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{ji} a_j(X(t)) + \sum_{j=1}^{M} \nu_{ji} \sqrt{a_j(X(t))}\Gamma_j(t), \tag{2.104}$$

where all parameters and variables are defined as the same as those in (2.56). $\Gamma_j(t)$ are temporally uncorrelated, statistically independent Gaussian white noise signals, and are formally defined by

$$\Gamma_j(t) = \lim_{dt\to\infty} \mathcal{N}(0, 1/\mathrm{d}t), \tag{2.105}$$

where $\mathcal{N}(m, \sigma^2)$ denotes the normal random variable with mean m and variance σ^2. Together with the properties of temporal and statistical independence, the definition implies

$$< \Gamma_j(t)\Gamma_{j'}(t') >= \delta_{jj'}\delta(t - t'), \tag{2.106}$$

where the first and second delta functions are Kronecker's function and Dirac's function, respectively, *i.e.*, $\delta_{jj'} = 1$ if $j = j'$, and 0 otherwise; $\delta(t - t') = 1$ if $t = t'$, and 0 otherwise. Next, we will occasionally use $\delta(j, j')$ to represent $\delta_{jj'}$.

Although the Langevin approach is approximate analysis due to continuous approximation of discrete state X, which loses validity when the number of molecules is small, it is often solved with much greater analytical ease than other representations, *e.g.*, the master equational approach. For example, on the basis of the transfer function around the feedback loop and the equivalent noise bandwidth of a system, a frequency domain technique for the analysis of intrinsic noise in negatively auto-regulated gene circuits has been proposed (Simpson Siegal-Gaskins 2003). Working with Langevin equations is beneficial, especially, when one can clearly find how the internal noise involved in biochemical reactions is related to the parameter values, the system size, and the state variables that evolve with time. Because internal noise is inherent in biochemical reactions and cannot be switched off, the Langevin equations have been extensively used in modeling intrinsic noise in cellular systems to study their constructive roles such as internal noise stochastic resonance (INSR) (Hou and Xin 2003) and coherence resonance (Steuer *et al.* 2003).

Let us take a biochemical clock model, which incorporates three species: mRNA (M), clock proteins in cytosol (P_C), and in nucleus (P_N), as an example (Hou and Xin 2003). The deterministic equations that describe the evolution of the three species are

$$\frac{\mathrm{d}[M]}{\mathrm{d}t} = \nu_s \frac{k_1^n}{k_1^n + [P_N]^n} - \nu_m \frac{[M]}{k_m + [M]} \triangleq a_1 - a_2, \tag{2.107}$$

$$\begin{aligned} \frac{\mathrm{d}[P_C]}{\mathrm{d}t} &= k_s[M] - \nu_d \frac{[P_C]}{k_d + [P_C]} - k_1[P_C] + k_2[P_N] \\ &\triangleq a_3 - a_4 - a_5 + a_6, \end{aligned} \tag{2.108}$$

$$\frac{\mathrm{d}[P_N]}{\mathrm{d}t} = k_1[P_C] - k_2[P_N] \triangleq a_5 - a_6. \tag{2.109}$$

According to (2.104), the Langevin equations for this system can be written directly as

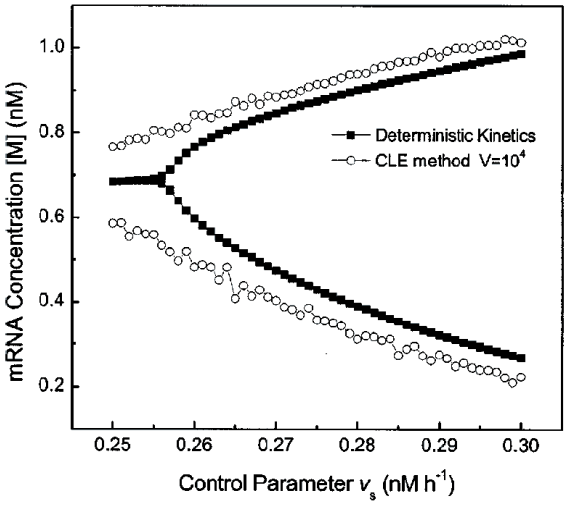

Figure 2.8 Bifurcation diagrams for the deterministic equations (2.107)–(2.109) and the chemical Langevin equations (CLE) (2.110)–(2.112) (from (Hou and Xin 2003))

$$\frac{\mathrm{d}[M]}{\mathrm{d}t} = (a_1 - a_2) + \frac{1}{\sqrt{V}}[\sqrt{a_1}\Gamma_1(t) - \sqrt{a_2}\Gamma_2(t)], \tag{2.110}$$

$$\frac{\mathrm{d}[P_C]}{\mathrm{d}t} = (a_3 - a_4 - a_5 + a_6) + \frac{1}{\sqrt{V}}[\sqrt{a_3}\Gamma_3(t) - \sqrt{a_4}\Gamma_4(t) - \sqrt{a_5}\Gamma_5(t) + \sqrt{a_6}\Gamma_6(t)], \tag{2.111}$$

$$\frac{\mathrm{d}[P_N]}{\mathrm{d}t} = (a_5 - a_6) + \frac{1}{\sqrt{V}}[\sqrt{a_5}\Gamma_5(t) - \sqrt{a_6}\Gamma_6(t)], \tag{2.112}$$

where $\Gamma_{i=1,\ldots,6}(t)$ are Gaussian white noise signals with $\langle\Gamma_i(t)\rangle = 0$ and $\langle\Gamma_i(t)\Gamma_j(t')\rangle = \delta_{ij}\delta(t-t')$, and V is the system size, *e.g.*, the cell volume. When removing the second terms in the brackets at the right side of (2.110)–(2.112), the equations are equivalent to the deterministic ones (2.107)–(2.109). Therefore, the second terms actually denote the internal noise. It is clear that the magnitude of the internal noise scales as $1/\sqrt{V}$ and also depends not only on the control parameters but also on system states, *i.e.*, the concentrations of $[M]$, $[P_C]$, and $[P_N]$.

Due to the explicit expression of noise in the Langevin equations, the constructive roles of intrinsic and extrinsic noise can be studied relatively easily. For example, as shown in Figure 2.8, when the system size is very large and the deterministic kinetics applies, the system would not sustain oscillations if the control parameter is less than the threshold. On the other hand, when the system size is small, stochastic oscillations occur in the Langevin equations for such a parameter region. Such stochastic oscillations are distinct from random

noise in that there is a clear peak in their power spectra. These oscillations are induced by the internal noise. In addition, the best performance at an optimal noise level demonstrates the occurrence of INSR (Hou and Xin 2003).

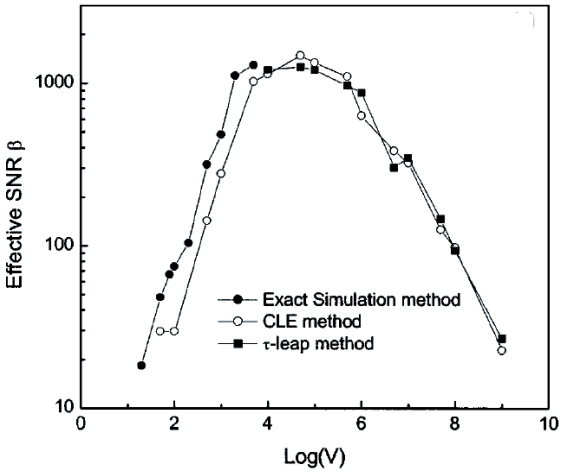

Figure 2.9 Dependence of the effective SNR on the system size at $\nu_s = 0.25$ (from (Hou and Xin 2003))

To measure the relative performance of the stochastic oscillations quantitatively, signal-to-noise ratio (SNR) can be defined, *i.e.*, $\beta = S/N$, where S and N are signal and noise strength, respectively. The dependence of SNR on the system size V is plotted in Figure 2.9. There is a clear maximum for system size $V \sim 10^4$, which demonstrates the existence of a resonance region. Since this resonance effect is purely induced by the internal noise, it is simply called INSR. See (Hou and Xin 2003) for an algorithm on how to calculate the effective SNR.

Recently, increasing amount of experimental and theoretical evidence indicates that noise can play a very important role in cellular systems. For example, noise-based switches and amplifiers were studied for gene expression in a single network derived from the bacteriophage λ (Hasty *et al.* 2000). Fluctuation-enhanced sensitivity of intracellular regulation in a single cell has also been reported (Paulsson *et al.* 2000). Internal noise can induce circadian oscillations, and the performance of the noise-induced circadian oscillation reaches a maximum with variation of the internal noise level, indicating the occurrence of INSR (Hou and Xin 2003, Li and Lang 2008). When combined with delays, a noisy system may keep oscillations for parameter values such that its corresponding deterministic one reaches a steady-state level (Bratsun *et al.* 2005, Lewis 2003). Noise-induced collective behavior was also discovered

for multicellular systems (Chen *et al.* 2005, Zhou *et al.* 2005). All these phenomena show that noise can induce potential richness to cellular dynamics.

2.5.5 Fokker–Planck Equations

Fokker–Planck equations are a special type of master equation, and are often used as an approximation to the actual master equations or as a model for more general Markov processes. By the Taylor expansion of $a_j(X(t) - \nu_j)P(X(t) - \nu_j)$ to order two in (2.56), we obtain the Fokker–Plank equation for the general molecular network as follows:

$$\frac{\partial P(X;t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial X_i} \left[\left(\sum_{j=1}^{M} \nu_{ji} a_j(X) \right) P(X;t) \right] + \frac{1}{2} \sum_{i=1}^{N} \frac{\partial^2}{\partial X_i \partial X_j} \left[\left(\sum_{j=1}^{M} \nu_{ji}^2 a_j(X) \right) P(X;t) \right], \quad (2.113)$$

where all parameters and variables are defined in (2.56). This equation can be rewritten in the following more compact form:

$$\frac{\partial P(X;t)}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial X_i} \left[A_i(X) P(X;t) \right] + \frac{1}{2} \sum_{i=1}^{N} \frac{\partial^2}{\partial X_i \partial X_j} \left[D_i(X) P(X;t) \right], \quad (2.114)$$

or in a vector form

$$\frac{\partial P(X;t)}{\partial t} = -\frac{\partial}{\partial X} \left[A(X) P(X;t) \right] + \frac{1}{2} \frac{\partial^2}{\partial X^2} \left[B(X) P(X;t) \right], \quad (2.115)$$

where

$$A(X) = (A_1, ..., A_N), \quad (2.116)$$

$$D(X) = (D_1, ..., D_N), \quad (2.117)$$

$$B(X) = [B_{ij}]_{N \times M}, \quad (2.118)$$

$$A_i(X) = \sum_{j=1}^{M} \nu_{ji} a_j(X), \quad (2.119)$$

$$B_{ij}(X) = \nu_{ji} a_j^{1/2}(X), \quad (2.120)$$

$$D_i(X) = \sum_{j=1}^{M} B_{ij}^2(X). \quad (2.121)$$

Note that the master equation is expanded with respect to variables X and discrete jumps ν_{ji}. If the variables are transformed as $x = X/V$ with the discrete jumps as ν_{ji}/V, where V is the system size (*e.g.*, the cellular volume), the master equation can be expanded with respect to x and ν_{ji}/V by the Taylor series, and such an approximation is called the Kramers–Moyal expansion. The first-order term of the Taylor or Kramers–Moyal expansion for the master equation is the deterministic kinetics of the molecular network, while the second-order term represents the Langevin dynamics.

It is clear that the Fokker–Planck equation (2.113) is a continuous approximation of discrete state X to the master equation (2.56) and can be proved to be equivalent to the Langevin equations (2.104). The Fokker–Planck equation is beneficial in the sense that some theoretical analysis can be conducted. For example, the equilibrium probability distribution can be obtained from (2.113) with one variable as follows:

$$P_{\text{eq}}(X) = \frac{\text{C}}{B(X)} \exp\left[2\int_0^X \frac{A(X')}{B(X')}\text{d}X'\right] \quad \text{and} \quad \frac{\partial}{\partial t}\langle X\rangle = \langle A_i(X)\rangle, \quad (2.122)$$

where C is a constant, and $\langle\cdot\rangle$ is the mean over all X. Note that P_{eq} is the equilibrium probability distribution in the completely stochastic case, not an equilibrium in the deterministic one. On the basis of the equilibrium probability distribution, the equilibrium mean concentration or number can be obtained.

The Fokker–Planck equation is an approximation to but similar to the master equation in that it describes the evolution of probability distribution of the state $X(t)$. One main difference between them is with regard to the presentation of the species. In the Fokker–Planck equation, the description is continuous, while in the master equation the representation is discrete. If the biochemical system contains only a few molecules, the discrete representation is more accurate than the continuous one.

As an example, consider the effect of a randomly varying external field or environment on the biochemical reactions. We will show how to transform a Langevin equation to an equivalent Fokker–Planck equation, or transform the master equation approximately to a Langevin equation. Reconsider the example (2.102) with additive noise $\Gamma(t)$, which is a rapidly fluctuating random term with zero mean $\langle\Gamma(t)\rangle = 0$ and is δ-correlated, *i.e.*, $\langle\Gamma(t)\Gamma(t')\rangle = D\delta(t-t')$ with D proportional to the strength of the perturbation.

Introducing the probability distribution $P(X;t)$, *i.e.*, the probability of the system with a state of concentration X at time t, its corresponding Fokker–Planck equation for $P(X;t)$ can be constructed as follows:

$$\begin{aligned}\frac{\partial P(X;t)}{\partial t} &= -\frac{\partial(f(X)P(X;t))}{\partial X} + \frac{D}{2}\frac{\partial^2 P(X;t)}{\partial X^2} \\ &= -\frac{\partial J(X;t)}{\partial X},\end{aligned} \quad (2.123)$$

where $J(X;t)$ is the following probability flux

$$J(X;t) = f(X)P(X;t) - \frac{D}{2}\frac{\partial P(X;t)}{\partial X}. \tag{2.124}$$

Compared with (2.115), in (2.124), clearly, $A(X) = f(X)$ and $B(X) = D$. At the equilibrium distribution, we have $J(X;t) = 0$. Then, integrating (2.124) over X, the equilibrium distribution for one-dimensional X is therefore given analytically as

$$P_{\rm eq}(X) = Ke^{-2/D\phi(X)} \tag{2.125}$$

with

$$\frac{\partial\phi(X)}{\partial X} = -f(X), \tag{2.126}$$

i.e., $\phi(X)$ can be viewed as an energy landscape, where K is a normalization constant determined by considering the integral of $P_{eq}(X)$ over all X to be unity (Hasty *et al.* 2000). Using (2.125), the equilibrium mean value is given by

$$\langle X\rangle_{\rm eq} = \int_0^\infty XP_{eq}(X){\rm d}X. \tag{2.127}$$

For multiplicative noise signals described by (2.103), the equilibrium probability distribution is obtained by transforming (2.103) to an equivalent Fokker–Planck equation

$$\frac{\partial P(X;t)}{\partial t} = -\frac{\partial(h(X) + \frac{D}{2}g(X)g'(X))P(X;t)}{\partial X} + \frac{D}{2}\frac{\partial^2 g^2(X)P(X;t)}{\partial X^2}, \tag{2.128}$$

where the prime denotes the derivative of $g(X)$ with respect to X. Compared with (2.115), $A(X) = h(X) + Dg(X)g'(X)/2$ and $B(X) = Dg^2(X)$. The equilibrium probability distribution for one-dimensional X can be similarly obtained as follows:

$$P_{\rm eq}(X) = Le^{-2/D\phi_m(X)} \tag{2.129}$$

with

$$\frac{\partial\phi_m(X)}{\partial X} = -\left(h(X) + \frac{D}{2}g(X)g'(X)\right), \tag{2.130}$$

where the function $\phi_m(X)$ can also be viewed as a potential (Hasty *et al.* 2000), and L is a normalization constant.

Clearly, from the master equation (2.56), we can derive the corresponding Langevin equations (2.104) and Fokker–Planck equation (2.115) for a general molecular network. Note that the Fokker–Planck equation (2.115) and the Langevin equations (2.104) are equivalent.

2.5.6 Cumulant Equations

It can be proven that when all state variables X follow Gaussian distributions, the Langevin equation (2.104) or the Fokker–Planck equation (2.113) can be equivalently expressed by the first and second cumulant evolution equations, which means that we can actually examine the dynamics by deterministic cumulant equations instead of the complicated stochastic dynamics such as the master equation and the Langevin equations. Next, we first describe a procedure to derive cumulant equations up to any order for a general cellular system, and then describe a closed-form expression of cumulant equations for the system with all variables following Gaussian distributions.

From the master equation (2.56) for a general molecular network, define

$$\bar{K}_i(X(t)) = \sum_{j=1}^{M} \nu_{ji} a_j(X(t)) \tag{2.131}$$

$$\bar{K}_{ik}(X(t)) = \sum_{j=1}^{M} \nu_{ji}\nu_{jk} a_j(X(t)) \tag{2.132}$$

for $i = 1, ..., N$ and $k = 1, ..., N$. Let the concentration of $X(t)$ be $x(t)$, *i.e.*, $x(t) = X(t)/V$, where V is the system size or the individual cellular volume. By using the concentration and letting

$$f_i(x(t)) = \bar{K}_i(Vx(t))/V, \quad K_{ik}(x(t)) = \bar{K}_{ik}(Vx(t))/V, \tag{2.133}$$

the Langevin equations (2.104) with molecular numbers as states can be rewritten into the following form with concentrations as states (Van Kampen 1992, Chen *et al.* 2005):

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = f_i(x(t)) + \xi_i(t), \tag{2.134}$$

where the vector ξ_i are Gaussian white noise signals with zero means and the intracellular covariances $K_{ik}(x(t))$, *i.e.*, $\langle \xi_i(t) \rangle = 0$ and $\langle \xi_i(t)\xi_k(t') \rangle = K_{ik}(x(t))\delta(t - t')$, and K_{ik} is an $N \times N$ matrix. $\langle \cdot \rangle$ is the means, or implementation of integration over the probability distribution, *i.e.*,

$$\langle f(x) \rangle = \int_x f(x) P(Vx; t) \mathrm{d}x, \tag{2.135}$$

where $X = Vx$. Clearly, Langevin equations (2.104) and (2.134) are equivalent but with different variables, and both are derived from (2.56).

The first and second cumulants for any two random variables x_i and x_j are actually their means and covariances, *i.e.*, $\langle x_i \rangle$, $\langle x_j \rangle$, and $\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$. Letting

$$g(x(t), s) = \prod_{i=1}^{N} x_i^{s_i}(t), \tag{2.136}$$

then for each integer-valued vector $s = (s_1, ..., s_N)$, the moment evolution equations corresponding to (2.134) are given as follows (Kawai *et al.* 2004, Wojtkiewicz *et al.* 1996),

$$\frac{\mathrm{d}\langle g(x(t), s)\rangle}{\mathrm{d}t} = \left\langle \sum_{i=1}^{N} f_i \frac{\partial g}{\partial x_i} \right\rangle + \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \left\langle K_{ik} \frac{\partial^2 g}{\partial x_i \partial x_k} \right\rangle, \tag{2.137}$$

which can be directly used to derive the cumulant evolution equations. Therefore, from (2.134) and (2.137), we can derive the cumulant equations up to any order for a general molecular network which, however, may not have a closed-form.

On the other hand, if each variable x_i in a system follows a Gaussian distribution, we can show that it can be expressed exactly by the first- and second-order cumulant equations in a closed-form. For such a system, all odd central moments vanish and any even central moment can be expressed as products of the second central moments. For instance,

$$\langle\, x_i x_j x_k x_l \rangle_c = \langle x_i x_j \rangle_c \langle x_k x_l \rangle_c + \langle x_i x_k \rangle_c \langle x_j x_l \rangle_c + \langle x_i x_l \rangle_c \langle x_j x_k \rangle_c, \tag{2.138}$$

$$\begin{aligned} \langle\, x_i x_i x_i x_j x_j x_k \rangle_c &= 6\langle x_i x_i \rangle_c \langle x_i x_j \rangle_c \langle x_j x_k \rangle_c + 6\langle x_i x_j \rangle_c^2 \langle x_i x_k \rangle_c \\ &\quad + 3\langle x_i x_i \rangle_c \langle x_i x_k \rangle_c \langle x_j x_j \rangle_c, \end{aligned} \tag{2.139}$$

where all moments are central moments (Chen *et al.* 2005), *e.g.*, $\langle x_i x_j \rangle_c = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$. Note that cumulants are identical to central moments for the first, second, and third orders, and any differentiable function $f(x)$ can be expanded around $\langle x \rangle$ by central moments, *i.e.*, for the Gaussian distribution of X, since all odd central moments are zero, $f(x)$ can be expanded as

$$\langle f(x) \rangle = f(\langle x \rangle) + \frac{1}{2!} \frac{\partial^2 f(\langle x \rangle)}{\partial x^2} \langle xx \rangle_c + \frac{1}{4!} \frac{\partial^4 f(\langle x \rangle)}{\partial x^4} \langle xxxx \rangle_c + \cdots. \tag{2.140}$$

The cumulant evolution equations of (2.134) with a closed-form expression can be derived in terms of the first-order cumulant m_i and the second-order cumulant c_{ik} as follows:

$$\frac{\mathrm{d}m_i(t)}{\mathrm{d}t} = F_i(m(t), c(t)), \tag{2.141}$$

$$\frac{\mathrm{d}c_{ik}(t)}{\mathrm{d}t} = G_{ik}(m(t), c(t)), \tag{2.142}$$

with

$$F_i(m(t), c(t)) = \langle f_i(x(t)) \rangle \tag{2.143}$$

$$\begin{aligned} G_{ik}(m(t), c(t)) &= \langle (x_i(t) - m_i(t)) f_k(x(t)) \rangle \\ &\quad + \langle (x_k(t) - m_k(t)) f_i(x(t)) \rangle + \langle K_{ik}(x(t)) \rangle, \end{aligned} \tag{2.144}$$

where $i, k = 1, ..., N$. In other words, a molecular network can be expressed exactly by the first-order and second-order cumulant equations. Here, the first cumulants or means of x are $m = (m_1, ..., m_N)$ with each element $m_i = \langle x_i \rangle$, and the second cumulants or covariances of x are $c = [c_{ik}]_{N \times N}$ with each element $c_{ik} = \langle (x_i - m_i)(x_k - m_k) \rangle$.

(2.141) is obtained by directly integrating (2.134) over all x_i, whereas (2.142) is derived by integrating $\mathrm{d}(x_i - m_i)(x_k - m_k)/\mathrm{d}t$ over all x_i and x_k with the substitution of (2.134), m_i, and m_k. All F_i and G_{ik} are the results of those integrating implementations. The vector m clearly has N elements. On the other hand, the non-zero elements of the covariance matrix c are at most $N(N+1)/2$, but more than N because any two molecules in each cell are not generally dependent on each other. A detailed example, which shows how to derive the explicit expression for the Langevin equations (2.134) and cumulant equations (2.141)–(2.142) for a gene regulatory network can be found in (Chen *et al.* 2005).

The system (2.141)–(2.142) is a closed-form expression in m and c, and hence can be viewed as a deterministic representation for a cellular system although there are high-order statistic terms, *i.e.*, covariances c_{ij} in addition to the means m_i. By examining deterministic cumulant equations, we can approximately determine the qualitative dynamics of the original stochastic system described by master or Langevin equations, *e.g.*, the stochastic stability and the stochastic synchronization from the viewpoint of probability distribution. For instance, if each x_i follows a Gaussian distribution, the switching dynamics, oscillatory behavior or synchrony in (2.141)–(2.142) correspond to those in (2.134), because the original stochastic dynamics of x can be entirely reconstructed from the deterministic means m and covariances c (Chen *et al.* 2005).

The fundamental importance of intrinsic and extrinsic noises within molecular networks has been appreciated within both mathematical and biological communities with increasing interest in recent years. These studies include techniques to predict (Mettetal *et al.* 2006), estimate (Orrell *et al.* 2005), and control (Orrell and Bolouri 2004) stochastic noise, effects of negative and positive feedback on intrinsic and external noise (Yi 2004), noise-induced qualitative change in dynamics (Hasty *et al.* 2000, Chen *et al.* 2005, J. Wang *et al.* 2007), enhancement of cellular memory by reducing stochastic transitions (Acar *et al.* 2005), robust properties with respect to noise (Gonze *et al.* 2002a), and some other stochastic cellular models to quantitatively or qualitatively investigate different roles played by noise (Tian and Burrage 2006, Tsimring *et al.* 2006). Elucidating constructive roles of stochastic noise and developing new techniques to deal with these stochastic fluctuations will remain an important and attractive subject of active research.

2.6 Deterministic Representation

There are two mathematical forms to model dynamical molecular networks, generally, *i.e.*, one is a stochastic formulation that explicitly includes the discrete and probabilistic change in reactant molecule numbers as reactions occur and the other is a deterministic formulation with reactant concentrations varying continuously in time and governed by a system of rate equations.

The advantage of the deterministic representation is that qualitative behavior can be obtained relatively easily by using theoretical methods of complex networks, nonlinear dynamics, and control. Therefore, with the further development of sophisticated analytical and computational techniques, deterministic representation is expected to not only be a powerful tool to provide testable predictions and foundations for designing synthetic molecular networks and controlling cellular processes but also have great potential for biotechnological and therapeutic applications.

2.6.1 Basic Kinetics

In the deterministic formulation, a cellular system or molecular network is considered to be a series of elementary biochemical reactions, whose kinetics can be described by rate equations according to the mass action law.

Consider the simplest monomolecular reaction, which is called the first-order reaction due to one reactant:

$$A \xrightarrow{k} A'. \tag{2.145}$$

It states that the species A is converted to species A' with a rate constant k. As the reaction occurs, $[A]$ decreases and $[A']$ increases. In this book, $[X]$ represents the concentration of species X, but we also directly use X to represent its concentration if no confusion arises. Many biochemical reactions such as transcription from DNA to an mRNA, translation from an mRNA to a protein, and transportation of a protein from the cytoplasm into the nucleus can be written in such a form. The reaction rate depends not on the products but only on the reactants and is a positive quantity identical for all participating species in an elementary reaction. Particularly, for the reaction (2.145), the rate equation has the form

$$r = \frac{\mathrm{d}[A']}{\mathrm{d}t} = -\frac{\mathrm{d}[A]}{\mathrm{d}t} = k[A]. \tag{2.146}$$

Bimolecular reactions probably are the most common reactions occurring in cellular systems. The two molecules can be the same or different. If the reactants are two different species, the reaction takes the following form, which is called the second-order reaction due to the presence of two reactants:

$$A + B \xrightarrow{k} AB. \tag{2.147}$$

Many biochemical reactions, *e.g.*, binding of a TF to DNA, formation of a heterodimer through binding of two different proteins, and conversion of a substrate into a complex catalyzed by an enzyme, take such a form. The reaction rate is

$$r = \frac{\mathrm{d}[AB]}{\mathrm{d}t} = -\frac{\mathrm{d}[A]}{\mathrm{d}t} = -\frac{\mathrm{d}[B]}{\mathrm{d}t} = k[A][B]. \tag{2.148}$$

When both the reactants are the same species, the reaction becomes

$$2A \xrightarrow{k} A_2. \tag{2.149}$$

Formation of a homodimer through binding of two monomers and formation of a tetramer through binding of two identical homodimers are such types. The reaction rate can be written as

$$r = \frac{\mathrm{d}[A_2]}{\mathrm{d}t} = -\frac{1}{2}\frac{\mathrm{d}[A]}{\mathrm{d}t} = k[A]^2. \tag{2.150}$$

The prefactor $1/2$ ensures that one gets the same rate for the reactants and products.

DNA is a relatively stable molecule; on the other hand, mRNAs and proteins can be constantly degraded by cellular machinery and recycled. The degradation of mRNAs and proteins takes the simple form which is also a first-order reaction:

$$A \xrightarrow{k} 0, \tag{2.151}$$

where A can be an mRNA or a protein. The reaction rate can be written as

$$r = -\frac{\mathrm{d}[A]}{\mathrm{d}t} = k[A]. \tag{2.152}$$

Besides monomolecular and bimolecular reactions, there are some other complex reactions such as reversible reactions, parallel reactions, and consecutive reactions, each of which can be decomposed into multiple elementary reactions. Of course, there are also some other types of reactions which occur rarely such as trimolecular reactions. We now develop formalisms for a general biochemical reaction. A general elementary biochemical reaction takes the form (see (2.1))

$$r_1R_1 + r_2R_2 + \cdots r_mR_m \xrightarrow{k} p_1P_1 + p_2P_2 + \cdots + p_nP_n, \tag{2.153}$$

where m is the number of reactants, and n is the number of products. With m reactants, (2.153) is the $\sum_{i=1}^{m} r_i$-order reaction due to $\sum_{i=1}^{m} r_i$ reactants. R_i is the ith reactant and P_j is the jth product. r_i and p_j are the numbers of reactant R_i consumed and product P_j produced in a reaction step, respectively. The coefficients r_i and p_j are known as stoichiometries. The reaction rate based on the mass action law is

$$r = -\frac{1}{r_i}\frac{\mathrm{d}[R_i]}{\mathrm{d}t} = \frac{1}{p_j}\frac{\mathrm{d}[P_j]}{\mathrm{d}t} = k\prod_{l=1}^{m}[R_l]^{r_l}, \tag{2.154}$$

where $i = 1, ..., m$ and $j = 1, ..., n$.

When more than one reaction are considered, the differential equations for a certain chemical is simply the sum of the contribution of each reaction. For example, based on rate equations, the reactions

$$A + B \overset{k}{\rightharpoonup} AB \tag{2.155}$$

$$AB \overset{k_{-1}}{\rightharpoonup} A + B \tag{2.156}$$

lead to a system of differential equations

$$\frac{\mathrm{d}[AB]}{\mathrm{d}t} = k_1[A][B] - k_{-1}[AB], \tag{2.157}$$

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = \frac{\mathrm{d}[B]}{\mathrm{d}t} = k_{-1}[AB] - k_1[A][B]. \tag{2.158}$$

The transformation is simply the sum of the rates of the two reactions.

Generally, a cellular system consists of a network of coupled biochemical reactions. These reactions may be types of transcription, translation, dimerization, protein or mRNA degradation, enzyme-catalyzed reactions, transportation, and transduction. Such reactions constitute various metabolic, genetic, and signaling networks. Moreover, one species may participate in multiple reactions, and one species can be a reactant in one reaction and a product in another one. Thus, the change of one species is actually the sum of changes in all reactions that it participates in.

2.6.2 Deterministic Representation of a General Molecular System

Recalling the general molecular network represented by master equation (2.56), we can also derive the deterministic representation in the form of differential equations with the numbers of molecules as variables by simply eliminating noise terms of (2.104) as

$$\frac{\mathrm{d}X_i(t)}{\mathrm{d}t} = \sum_{j=1}^{M}\nu_{ji}a_j(X(t)), \tag{2.159}$$

where all variables and parameters are defined as those in (2.104). The same expression but with the concentrations of molecules as variables can be obtained from (2.134) by eliminating $\xi_i(t)$.

Actually, the deterministic representation (2.159) for a general molecular network is identical to (2.141) when $X_i/v = m_i$ and all $c_{ij} = 0$. In other words, when we ignore the stochastic noise (or let the means represent whole dynamical states), (2.159) approximately captures the main dynamical behavior of the general molecular network described by (2.56).

2.6.3 Michaelis–Menten and Hill Equations

By exploiting multiple time scales, *e.g.*, fast–slow dynamics, the reaction (2.159) can be approximately expressed in a simple form. A commonly used reaction model for enzymatic reactions is the Michaelis–Menten (MM) equation, which is an approximation of the original dynamics. In this reaction mechanism it is assumed that the enzyme is neither consumed nor produced; therefore, the total concentration of the enzyme remains constant. It only interacts directly with the substrate to form an enzyme–substrate complex, which leads to synthesis of the product and release of the enzyme:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \overset{k_2}{\rightharpoonup} E + P, \tag{2.160}$$

where E, S, and P are enzyme, substrate, and product, respectively. The rate equations are now a system of differential equations:

$$\frac{\mathrm{d}[S]}{\mathrm{d}t} = -k_1[E][S] + k_{-1}[ES], \tag{2.161}$$

$$\frac{\mathrm{d}[E]}{\mathrm{d}t} = -k_1[E][S] + (k_{-1} + k_2)[ES], \tag{2.162}$$

$$\frac{\mathrm{d}[ES]}{\mathrm{d}t} = k_1[E][S] - (k_{-1} + k_2)[ES], \tag{2.163}$$

$$\frac{\mathrm{d}[P]}{\mathrm{d}t} = k_2[ES]. \tag{2.164}$$

The quasi-steady-state assumption for the enzyme–substrate complex ES due to the fast reaction, *i.e.*, $\mathrm{d}[ES]/\mathrm{d}t = 0$, leads to

$$[ES] = \frac{[E][S]}{K_M} \tag{2.165}$$

with the MM constant

$$K_M = \frac{k_{-1} + k_2}{k_1}. \tag{2.166}$$

Combining the quasi-steady-state complex concentration and the conservation law for the enzyme $[E]_\mathrm{T} = [E] + [ES]$ with the total enzyme concentration $[E]_\mathrm{T}$, we obtain

$$[ES] = \frac{[E]_\mathrm{T}[S]}{K_M + [S]} \tag{2.167}$$

for the enzyme. This leads to the well-known MM equation

$$\frac{\mathrm{d}[P]}{\mathrm{d}t} = \frac{V_{\max}[S]}{K_M + [S]}, \tag{2.168}$$

where $V_{\max} = k_2[E]_{\mathrm{T}}$ is the maximum reaction rate, and $[E]_{\mathrm{T}}$ is assumed to be constant.

In a molecular network, when there is no cooperative interactions among molecules such as dimerization, binding of TFs and their cofactors, and binding of TFs to an operator site of a promoter on DNA, a cellular process (*e.g.*, gene regulation) can be described approximately by the MM type of equations. Generally, gene expression can be regulated at multiple levels involving interactions among inducers (or cofactors), repressors, activators, and operator sites (or binding sites) on DNA. On one hand, a repressor (or an activator) protein, which is called a TF, binds to an operator site at the beginning of a gene to prevent (or promote) RNAP attaching to the DNA to synthesize mRNA. On the other hand, an inducer which is called a cofactor, binds to a repressor (or activator), causing it to change shape and preventing (or enhancing) its binding to the DNA strand. Therefore, the interactions among TFs, cofactors, and operator sites lead to two categories of transcriptional regulation: activation and repression.

Take the binding of a repressor protein to an inducer as an example. The repressor protein P binds to a small inducer I to form a complex PI by

$$P + I \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} PI. \tag{2.169}$$

The repressor is therefore found in either free (P) or bound (PI) form. Assume that the number of P is relatively stable in contrast to the small inducer I, *i.e.*, the conservation law states that the total concentration of the repressor protein remains constant of $[P]_{\mathrm{T}}$, *i.e.*,

$$[P]_{\mathrm{T}} = [P] + [PI]. \tag{2.170}$$

The kinetic rate equation is

$$\frac{\mathrm{d}[PI]}{\mathrm{d}t} = k_1[P][I] - k_{-1}[PI]. \tag{2.171}$$

A system is said to be at equilibrium when its state ceases to change. Assume that the system has reached its equilibrium, that is, there are no further net changes. Thus,

$$\frac{\mathrm{d}[PI]}{\mathrm{d}t} = k_1[P][I] - k_{-1}[PI] = 0, \tag{2.172}$$

which leads to

$$K_{\mathrm{eq}}[PI] = [P][I], \tag{2.173}$$

where $K_{\mathrm{eq}} = k_{-1}/k_1$ is called the equilibrium constant. Combining the conservation of the total repressor concentration, we derive the MM equation as follows:

$$[PI] = \frac{[P]_{\mathrm{T}}[I]}{[I] + K_{\mathrm{eq}}}. \tag{2.174}$$

The MM function has three notable features (Alon 2006):

1. it reaches saturation at high $[I]$;
2. it has a regime where $[PI]$ increases linearly with I when $[I] << K_{\mathrm{eq}}$;
3. the fraction of bound protein reaches 50% when $[I] = K_{\mathrm{eq}}$.

Another commonly used function is the Hill type of equation, which can be viewed as an extension of the MM function. As well known, many molecules or TFs have several identical subunits or are composed of several identical proteins, *e.g.*, in the form of dimers or tetramers. The Hill function is often used to describe cooperative interactions among such molecules and can be derived as follows. Assume that the repressor protein P has n active binding sites. When n identical inducer monomers I bind to the protein P, the protein P can be either bound to n molecules of I, described by complex PI_n, or unbound, denoted by P. The reaction takes the form

$$P + nI \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} PI_n. \tag{2.175}$$

Assume that the total concentration of the bound and unbound P is P_{T}, which remains constant. The conservation law takes the form

$$[P]_{\mathrm{T}} = [P] + [PI_n]. \tag{2.176}$$

The kinetic rate equation for such a reaction is

$$\frac{\mathrm{d}[PI_n]}{\mathrm{d}t} = k_1[P][I]^n - k_{-1}[PI_n], \tag{2.177}$$

which reaches an equilibrium usually within milliseconds and therefore leads to a steady-state approximation, *i.e.*, $k_1[P][I]^n = k_{-1}[PI_n]$. Combining the conservation law (2.176), we derive the Hill equation as follows:

$$[PI_n] = \frac{[P]_{\mathrm{T}}[I]^n}{K_{\mathrm{eq}}^n + [I]^n} \triangleq f([I]), \tag{2.178}$$

where $K_{\mathrm{eq}}^n = k_{-1}/k_1$. Similarly, we can derive the MM function and the Hill function for binding of TF proteins and operator sites based on the conservation of operator sites. The parameter K_{eq} is termed the activation coefficient. Half-maximum binding $[P]_{\mathrm{T}}/2$ is reached at $[I] = K_{\mathrm{eq}}$ and a limiting value, *i.e.*, maximum binding $[P]_{\mathrm{T}}$, is reached at a sufficiently high level of $[I]$. This saturation of the Hill function is fundamentally due to physical restriction of binding sites of the promoter. The parameter n is termed as the Hill coefficient. The larger is n, the steeper of the Hill curve (Alon 2006). When $n = 1$, we obtain the MM equation.

The Hill equation (2.178) is a monotonically increasing function with respect to $[I]$; therefore, it is the case where I acts as an enhancer. For a repressor case, the Hill function takes the form

$$f([I]) = \frac{[P]_{\mathrm{T}}}{1 + ([I]/K_{\mathrm{eq}})^n}, \tag{2.179}$$

which is a decreasing function of I.

The MM and Hill equations are commonly used in modeling molecular networks. For example, when studying network topologies that can achieve adaption, *i.e.*, the ability to reset signaling systems after responding to stimuli, the MM functions are used directly to model enzymatic regulatory networks, as shown in Figure 2.10. It was shown that only two major core topologies emerge as robust adaption: a negative feedback loop with a buffering node and an incoherent feedforward loop with a proportioner node (Ma *et al.* 2009). Moreover, the use of Hill functions to model the enzymatic regulatory networks does not significantly alter their results (Ma *et al.* 2009). When modeling gene networks (or genetic networks), MM and Hill kinetics are generally used for transcriptional rates, and mass action kinetics are used for all other rates, *e.g.*, mRNA and protein degradations, translation, and complex formation and dissociation (Mirsky *et al.* 2009).

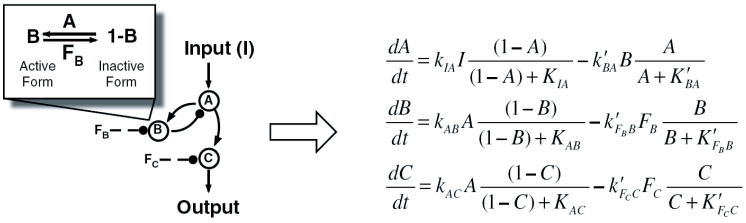

Figure 2.10 Illustration using MM functions to model one enzymatic regulatory network (from (Ma *et al.* 2009))

For transcriptional processes, the mathematical expression of the toggle switch (Gardner *et al.* 2000) is another example demonstrating how to use the MM and Hill equations to model a molecular network. Specifically, let us consider a case where each of the two proteins negatively regulates the synthesis of the other, as depicted in Figure 2.11 (Gardner *et al.* 2000). The dynamics of the toggle switch can be described by the dimensionless model

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\alpha_1}{1 + v^\beta} - u, \tag{2.180}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\alpha_2}{1 + u^\gamma} - v, \tag{2.181}$$

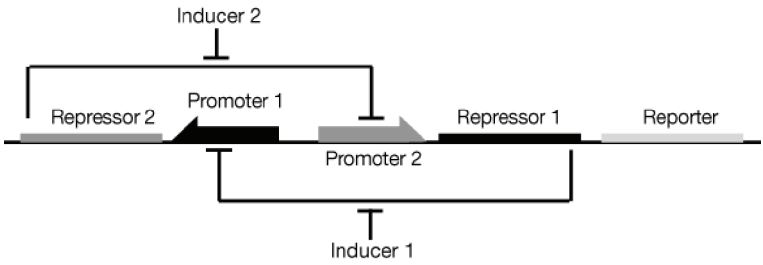

Figure 2.11 The toggle switch is composed of two co-repressive genes. The constitutive promoter 1 drives the expression of the *lacI* gene, which produces the lac repressor tetramer. The lac repressor tetramer binds the lac operator sites adjacent to the promoter 2, thereby blocking transcription of *cI*. The constitutive promoter 2 drives the expression of the *cI* gene, which produces the λ-repressor dimer. The λ-repressor dimer cooperatively binds to the operator sites within the promoter 1, which prevents transcription of *lacI* (from (Gardner *et al.* 2000))

where u is the concentration of the repressor 1; v is the concentration of the repressor 2; α_1 and α_2 are the effective rates of synthesis of the repressor 1 and repressor 2, respectively; β and γ are the cooperativity parameters of repression of the promoter 2 and promoter 1, respectively. The first terms on the right side of (2.180)–(2.181) represent production of the two repressors due to transcription. The system equations preserve the two most fundamental aspects of the network: mutual repression of constitutively transcribed genes and degradation/dilution of the repressors, described by the first and second terms, respectively, in each equation (Gardner *et al.* 2000).

2.6.4 Total Quasi-steady-state Approximation

Although the MM approximation is widely used to study enzyme kinetics, it is not applicable to the signaling cycle because the MM approximation assumes the much lower concentration of the enzyme than that of the substrate, which is not generally valid in signaling pathways. For example, substrates and enzymes of MAPK pathways are usually present at comparable concentrations in *Saccharomyces cerevisiae* and *Xenopus* oocyte cells (Ferrell 1996). Recently, the total quasi-steady-state approximation (tQSSA) which is valid more generally was proposed and applied to model biomolecular networks of coupled enzymatic reactions, including interconnection of phosphorylation (Ciliberto *et al.* 2007). The tQSSA for the irreversible (Tzafriri 2003) and reversible MM schemes (Tzafriri and Edelmana 2004) have been developed for the limit of low enzyme concentrations and also for high enzyme concentrations. Their reduced tQSSA representation accurately reproduces behavior predicted by detailed mass action kinetics models.

Consider the signaling cycle in signal transduction consisting of a substrate protein that can be in one of two states, either active A or inactive I (Gomez

et al. 2007). The signaling cycle can be modeled by two enzymatic reactions, which have the form of MM equations similar to (2.168):

$$I + E_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} IE_1 \overset{k_2}{\rightarrow} A + E_1, \tag{2.182}$$

$$A + E_2 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} AE_2 \overset{k_4}{\rightarrow} I + E_2, \tag{2.183}$$

where E_1 is the kinase and E_2 is the phosphatase.

Considering that the six chemical species, *i.e.*, I, E_1, IE_1, A, E_2, and AE_2, and three conservation relations, *i.e.*, that the kinase, the phosphatase, and the substrate protein, are conserved, we obtain a system of three variables according to the mass action law:

$$\frac{\mathrm{d}\bar{A}}{\mathrm{d}t} = k_2 IE_1 - k_4 AE_2, \tag{2.184}$$

$$\frac{\mathrm{d}IE_1}{\mathrm{d}t} = k_1[(\bar{S} - \bar{A} - IE_1)(\bar{E}_1 - IE_1) - K_1 IE_1], \tag{2.185}$$

$$\frac{\mathrm{d}AE_2}{\mathrm{d}t} = k_3[(\bar{A} - AE_2)(\bar{E}_2 - AE_2) - K_2 AE_2], \tag{2.186}$$

where

$$K_1 = \frac{k_2 + k_{-1}}{k_1} \quad \text{and} \quad K_2 = \frac{k_4 + k_{-3}}{k_3} \tag{2.187}$$

are the MM constants for the kinase and the phosphatase, respectively. $\bar{S}$ stands for the total concentration of the substrate protein in both active and inactive forms, *i.e.*, $\bar{S} = I + A$. X denotes the concentration of an unbound chemical species, and $\bar{X}$ denotes the total concentration of bound and unbound forms, *i.e.*, $\bar{A} = A + AE_2$, $\bar{E}_1 = E_1 + IE_1$, and $\bar{E}_2 = E_2 + AE_2$ (Gomez *et al.* 2007).

Assuming that the complexes IE_1 and AE_2 have faster dynamics than the active protein $\bar{A}$ and that they are always at equilibrium with respect to the active substrate protein, the following algebraic expressions are obtained:

$$IE_1 = \frac{K_1 + \bar{E}_1 + \bar{S} - \bar{A}}{2}(1 - \sqrt{1 - 4r_1}), \tag{2.188}$$

$$AE_2 = \frac{K_2 + \bar{E}_2 + \bar{A}}{2}(1 - \sqrt{1 - 4r_2}), \tag{2.189}$$

where

$$r_1 = \frac{\bar{E}_1(\bar{S} - \bar{A})}{(K_1 + \bar{E}_1 + \bar{S} - \bar{A})^2}, \tag{2.190}$$

$$r_2 = \frac{\bar{E}_2 \bar{A}}{(K_2 + \bar{E}_2 + \bar{A})^2}. \tag{2.191}$$

Approximating to the first order in r_1 and r_2 yields the expressions

$$IE_1 = \frac{\bar{E}_1(\bar{S} - \bar{A})}{(K_1 + \bar{E}_1 + \bar{S} - \bar{A})}, \tag{2.192}$$

$$AE_2 = \frac{\bar{E}_2\bar{A}}{(K_2 + \bar{E}_2 + \bar{A})}. \tag{2.193}$$

Inseting (2.192)–(2.193) into (2.184), we obtain the time evolution of the total active protein as follows:

$$\frac{\mathrm{d}\bar{A}}{\mathrm{d}t} = k_2 \frac{\bar{E}_1(\bar{S} - \bar{A})}{K_1 + \bar{E}_1 + \bar{S} - \bar{A}} - k_4 \frac{\bar{E}_2\bar{A}(t)}{K_2 + \bar{E}_2 + \bar{A}}. \tag{2.194}$$

2.6.5 Deriving Rate Equations

Now, we present a procedure for deriving rate equations of a cellular system. The procedure is summarized as follows:

1. Draw an illustration of all reactions to be considered. It contains all the related reactants and products.
2. Introduce the respective concentrations of substances as system states.
3. For each substance S, according to the mass action law, sum up the rates of all reactions that it participates in, *i.e.*, reactions which transform other substances to S and reactions which transform S to other substances.
4. Simplify the system or the network by utilizing the inherent separation of time scales, *i.e.*, by allowing fast equations to be at their respective equilibrium.
5. Further simplify the system by conservation laws, *e.g.*, total enzyme concentrations and the number of total binding sites of each promoter remain constant.

Consider the example of the toggle switch. To show how to obtain (2.180)–(2.181), we write the biochemical reactions governing the process in detail. Generally, biochemical reactions are divided into two categories: fast and slow reactions. The fast reactions are mainly phosphorylation, multimerization, and binding processes and have rate constants of order seconds, and are assumed to be in equilibrium with respect to the slow reactions. Letting R_1, P_1, R_2, P_2, R_1^γ, and R_2^β denote the repressor 1, promoter 1 (or binding site 1), repressor 2, promoter 2 (or binding site 2), repressor 1 multimer (or γ-mer), and repressor 2 multimer (or β-mer), respectively, in Figure 2.11, we can write the biochemical reactions as follows:

$$\gamma R_1 \overset{K_3}{\rightleftharpoons} R_1^{\gamma}, \tag{2.195}$$

$$\beta R_2 \overset{K_4}{\rightleftharpoons} R_2^{\beta}, \tag{2.196}$$

$$P_1 + R_1^{\gamma} \overset{K_1}{\rightleftharpoons} P_1 R_1^{\gamma}, \tag{2.197}$$

$$P_2 + R_2^{\beta} \overset{K_2}{\rightleftharpoons} P_2 R_2^{\beta}, \tag{2.198}$$

where $K_i = k_i/k_{-i}$ are equilibrium constants. All these reactions generally evolve considerably fast, *e.g.*, in terms of less than seconds, and reach equilibrium quickly, compared with the transcription and translation reactions, which usually require more than a few minutes.

On the other hand, the slow irreversible reactions, such as transcription of mRNAs, translation of proteins, and degradation of proteins and mRNAs, evolve on a time scale that is much slower than those of fast reactions (2.195)–(2.198). The slow reactions are mainly a synthesis of the two repressors and their degradation

$$P_2 \overset{k_1'}{\rightarrow} P_2 + nR_1, \tag{2.199}$$

$$P_2 R_2^{\beta} \overset{k_1''}{\rightarrow} P_2 R_2^{\beta} + nR_1, \tag{2.200}$$

$$P_1 \overset{k_2'}{\rightarrow} P_1 + mR_2, \tag{2.201}$$

$$P_1 R_1^{\gamma} \overset{k_2''}{\rightarrow} P_1 R_1^{\gamma} + mR_2, \tag{2.202}$$

$$R_1 \overset{d_1}{\rightarrow} 0, \tag{2.203}$$

$$R_2 \overset{d_2}{\rightarrow} 0, \tag{2.204}$$

where n and m denote the number of the two repressors per mRNA transcript, respectively. Here, for the sake of simplicity, assume that $k_1'' = k_2'' = 0$ due to the strong repression of R_1 and R_2.

There are also conservation conditions for the total binding sites of the two promoters, *i.e.*, $P_1 + P_1 R_1^{\gamma} = P_2 + P_2 R_2^{\beta} = P_{\mathrm{T}}$, where P_{T} is the concentration of the genes *cI* and *lac* and is assumed to be the same for the two genes. By setting all (2.195)–(2.198) to be at their respective equilibrium due to their fast reactions, the synthesis rates of the two repressors are

$$V_1 = k_1' n[P_2] = \frac{k_1' n[P_2][P_{\mathrm{T}}]}{[P_2] + [P_2 R_2^{\beta}]} = \frac{k_1' n[P_{\mathrm{T}}]}{1 + K_2 K_4 [R_2]^{\beta}}, \tag{2.205}$$

$$V_2 = k_2' m[P_1] = \frac{k_2' m[P_1][P_{\mathrm{T}}]}{[P_1] + [P_1 R_1^{\gamma}]} = \frac{k_2' m[P_{\mathrm{T}}]}{1 + K_1 K_3 [R_1]^{\gamma}}. \tag{2.206}$$

Assuming the same degradation rates $d_1 = d_2 = \delta$ leads to

$$\frac{\mathrm{d}[R_1]}{\mathrm{d}t} = \frac{k_1' n[P_\mathrm{T}]}{1 + K_2 K_4 [R_2]^\beta} - \delta[R_1], \tag{2.207}$$

$$\frac{\mathrm{d}[R_2]}{\mathrm{d}t} = \frac{k_2' m[P_\mathrm{T}]}{1 + K_1 K_3 [R_1]^\gamma} - \delta[R_2]. \tag{2.208}$$

We next eliminate some of the parameters by rescaling the repressors $[R_1]$, $[R_2]$, and time. The dimensionless variables are defined as follows:

$$u = (K_1 K_3)^{1/\gamma}[R_1], \tag{2.209}$$

$$v = (K_2 K_4)^{1/\beta}[R_2], \tag{2.210}$$

$$\tau = t\delta. \tag{2.211}$$

By substituting (2.209)–(2.211) into (2.207)–(2.208), we finally obtain (2.180)–(2.181) with cooperativity parameters β and γ, respectively, and effective rates

$$\alpha_1 = \frac{k_1' n[P_T]}{\delta(K_1 K_3)^\gamma}, \tag{2.212}$$

$$\alpha_2 = \frac{k_2' m[P_T]}{\delta(K_2 K_3)^\beta}. \tag{2.213}$$

Many other detailed examples, *e.g.*, autoregulation of λ repressor expression (Hasty *et al.* 2000) and a synthetic multicellular gene network (Chen *et al.* 2005), can also be found in related literature. Details on differential equation models will be discussed in the succeeding chapters. In addition to the above procedure, as indicated in (2.159), we can also directly derive deterministic rate equations from a general molecular network (2.56).

2.6.6 Modeling Transcription and Translation Processes

Transcription and translation are fundamental processes in cellular systems. In particular, transcription is a key process in gene regulation, which results in the synthesis of RNA under the direction of DNA, *i.e.*, transcribing a gene or a DNA nucleotide sequence into a RNA sequence or mRNA, which is enzymatically copied by RNAP and is regulated by various TFs. When two or more TFs bind to the promoter of a gene, similar biochemical reactions shown in the previous section for a transcription process can be derived but with higher-order terms. Generally, to model gene regulation, transcription and translation processes are considered as slow reactions, whereas other biochemical reactions are viewed as fast reactions, which are therefore all reduced to equilibrium based on the quasi-steady-state assumption. Consider the system of (2.9)–(2.16) with the conservation condition (2.18) shown in Figure 2.1 as an example. By the quasi-steady-state assumption for (2.9)–(2.12), *i.e.*, $\mathrm{d}[P \cdot D]/\mathrm{d}t = \mathrm{d}[Q \cdot D]/\mathrm{d}t = \mathrm{d}[P \cdot Q \cdot D]/\mathrm{d}t = 0$ with the conservation condition (2.18), we can derive the following equations or configurations:

$$[D] = \frac{n_x}{1 + K_1[P] + K_2[Q] + K_{PQ}}, \tag{2.214}$$

$$[PD] = K_1[P][D] = \frac{K_1 n_x[P]}{1 + K_1[P] + K_2[Q] + K_{PQ}}, \tag{2.215}$$

$$[QD] = K_2[Q][D] = \frac{K_2 n_x[Q]}{1 + K_1[P] + K_2[Q] + K_{PQ}}, \tag{2.216}$$

$$[PQD] = (K_3 + K_4)[P][Q][D] = \frac{(K_3 + K_4) n_x [P] [Q]}{1 + K_1[P] + K_2[Q] + K_{PQ}}, \tag{2.217}$$

where $K_1 = k_1/k_{-1}$, $K_2 = k_2/k_{-2}$, $K_3 = k_3 K_2/(k_{-3} + k_{-4})$, $K_4 = k_4 K_1/(k_{-3} + k_{-4})$, and $K_{PQ} = (K_3 + K_4)[P][Q]$.

Generally, once RNAP binds to the DNA strand, transcription is initiated to produce mRNA. Let $P(\text{RNAP}_i)$ be the probability that the RNAP is bound to DNA or gene i. Then for a molecular network with n genes, the probability has the following form for $i = 1, ..., n$

$$P(\text{RNAP}_i) = \sum_{c_i \in C} P(c_i) P(\text{RNAP}_i | c_i), \tag{2.218}$$

where C is the set of all possible configurations of TFs with RNAP on the DNA, $P(c_i)$ is the probability distribution over TF configurations on the DNA, and $P(\text{RNAP}_i|c_i)$ is the probability (or affinity) that the RNAP is bound given the constellation of TFs bound to the ith gene in configuration $c_i \in C$. Here, we assume that the gene expression level, *i.e.*, the concentration of mRNA is proportional to the probability that the RNAP occupies its regulatory sequence. That is, the transcription and translation processes for the ith ($i = 1, ..., n$) gene can be expressed as

$$\frac{\text{d}[\text{mRNA}_i]}{\text{d}t} = k_{Ri} P(\text{RNAP}_i) - d_{Ri} \, [\text{mRNA}_i], \tag{2.219}$$

$$\frac{\text{d}[\text{Protein}_i]}{\text{d}t} = k_{Pi} [\text{mRNA}_i] - d_{Pi} [\text{Protein}_i], \tag{2.220}$$

where k_{Ri} and k_{Pi} denote the synthesis rates of the mRNA and the protein, respectively; d_{Ri} and d_{Pi} represent the degradation rates of the mRNA and the protein, respectively. Equations (2.219)–(2.220) represent a gene regulatory network.

For the case of Figure 2.1 with one gene, there are eight configurations of TFs and RNAP, *i.e.*, Figure 2.1 (a)–(h) for c_i, *i.e.*, gene-i, but $P(\text{RNAP}|c_i)$ is non-zero only for the cases of Figure 2.1 (b), (e)–(g). Assume that the affinities or probabilities of RNAP bound to the gene-i are $P(\text{RNAP}|(b)) = w_D$, $P(\text{RNAP}|(e)) = w_{PD}$, and $P(\text{RNAP}|(f)) = w_{QD}, P(\text{RNAP}|(g)) = w_{PQD}$ for the configurations of Figure 2.1 (b), (e)–(g), respectively, where we drop the subscript i because there is only one gene. Therefore, according to (2.218), we have

$$\begin{aligned} P(\mathrm{RNAP}_i) &= P(b)P(\mathrm{RNAP}|(b)) + P(e)P(\mathrm{RNAP}|(e)) \\ &\quad + P(f)P(\mathrm{RNAP}|(f)) + P(g)P(\mathrm{RNAP}|(g)) \\ &= \frac{w_D n_x + w_{PD}K_1[P][D]}{1 + K_1[P] + K_2[Q] + (K_3 + K_4)[P][Q]} \\ &\quad + \frac{w_{QD}K_2[Q][D] + w_{PQD}(K_3 + K_4)[P][Q][D]}{1 + K_1[P] + K_2[Q] + (K_3 + K_4)[P][Q]}, \end{aligned} \tag{2.221}$$

where $P(b) = [D]$, $P(e) = [PD]$, $P(f) = [QD]$, and $P(g) = [PQD]$ are given in (2.214)–(2.217). Therefore, substituting (2.221) with (2.219), we can obtain the local gene regulatory network with (2.220) for gene-i. In the same way, constructing equations for all studied genes, we can derive the gene regulatory network of the system, where the interactions among proteins, *i.e.*, $\mathrm{Protein}_i$, can be obtained from the related protein interaction network or signal pathway.

According to thermodynamic equilibrium, we can also derive the equation for $P(\mathrm{RNAP}_i)$ in the same form. Specifically, we assume that the transcription and translation rates are much slower than other biochemical reactions such as binding and dimerization. Therefore, the system is in a state of thermodynamic equilibrium such that each configuration is achieved with a probability that is proportional to its probability under the Boltzmann distribution. If for the ith gene there are m_i TFs, among which the jth TF has s_{ij} binding sites, then

$$P(\mathrm{RNAP}_i) = \frac{\sum_{q_{i1}=0}^{s_{i1}} \cdots \sum_{q_{im_i}=0}^{s_{im_i}} w_1[\mathrm{TF}]_{\mathrm{Q}}}{1 + \sum_{q_{i1}=0}^{1} \sum_{q_{i1}=0}^{s_{i1}} \cdots \sum_{q_{im_i}=0}^{s_{im_i}} w_0[\mathrm{TF}]_{\mathrm{Q}}}, \tag{2.222}$$

where $w_1 = w(1, q_{i1}, ..., q_{im_i})$ and $w_0 = w(q_0, q_{i1}, ..., q_{im_i})$ are the parameters related to the thermodynamic coefficients (or affinities) on the system and $[\mathrm{TF}]_{\mathrm{Q}} = [\mathrm{TF}_{i1}]^{q_{i1}} \cdots [\mathrm{TF}_{im_i}]^{q_{im_i}}$. $[\mathrm{TF}_{ij}]$ means the concentration of the jth TF for gene i. Therefore, substituting (2.222) into (2.219), we have the nonlinear dynamical equations of the gene regulatory network (2.219)–(2.220). Clearly, (2.219)–(2.220) with (2.222) are a general representation of a transcription regulatory network (Chen *et al.* 2009), which is applicable to a large class of systems. For example, by fitting parameters w, q, k, and d, we can obtain the transcription regulatory relations for any molecular network.

As a special example, assume that all TFs independently bind at available distinct sites of a gene's promoter region without interacting with each other, *i.e.*, there is no cooperativity between TFs. Then, the molecular network with n genes can be mathematically expressed as

$$\begin{aligned} \frac{\mathrm{d}[\mathrm{mRNA}_i]}{\mathrm{d}t} &= d_{i0} - d_i[\mathrm{mRNA}_i] \\ &\quad + d_i\Big(1 - \prod_{j \in R_i^+} \rho([\mathrm{TF}_i], s_{ij}, \theta_{ij})\Big) \prod_{j \in R_i^-} \rho([\mathrm{TF}_i], s_{ij}, \theta_{ij}), \end{aligned} \tag{2.223}$$

and (2.220), where $\rho([\mathrm{TF}], s, \theta) = 1/(1 + \theta[\mathrm{TF}])^s$. Let R_i be the regulator or TF set of the ith gene. Then, R_i^+ is the set of activators in R_i and R_i^- is the set of inhibitors in R_i. Hence, $R_i = R_i^+ \cup R_i^-$. s_{ij} represents the number of binding sites for the jth TF of gene i, and θ_{ij} represents the affinity constant for the jth TF to the binding site of gene i. d_{io} is the transcription initiation rate of the ith gene, and d_i is the degradation rate of the ith gene. Due to the assumption of independence between TFs, there is a considerably smaller set of parameters in the above expression, compared with those in (2.222).

2.7 Hybrid Representation and Reducing Molecular Networks

A living organism is inherently a discrete and stochastic system due to the random birth and death of molecules in cells. Explicitly considering all variables and chemical reactions in a cell is unrealistic for a molecular network, in particular, for a gene regulatory network from modeling, analyzing, and computing viewpoints. However, in a cell, many different time scales characterize the gene regulatory processes, which can be exploited to reduce the complexity of the mathematical models. For instance, the transcription and translation processes generally evolve on a time scale that is much slower than those of phosphorylation, dimerization, and binding reactions. Moreover, in biological systems, there are many subsystems, such as gene regulatory networks, protein interaction networks, and metabolic networks, which can dynamically interact with each other but are relatively independent. In this section, we exploit such properties to simplify a complicated molecular network to a hybrid system or even a deterministic system by developing several models, which can be applied to the quantitative simulation of a large cellular system. Note that there are two different implications for hybrid systems, *i.e.*, one is a hybrid system with both discrete and continuous variables, and the other is a hybrid system with both stochastic and deterministic dynamics, or stochastic hybrid processes. Hybrid simplification approximates some or all discrete variables by continuous ones, and therefore can drastically reduce computational time by eliminating the simulated discrete events from a computational viewpoint because the computational complexity depends on the number of discrete jumps. Besides the hybrid simplification based on the central limit theorem, other simplification schemes, such as averaging approximation and stochastic quasi-steady-state approximation, can also be adopted to simplify the complexity of the complicated molecular networks.

2.7.1 Decomposition of Biomolecular Networks

Consider a general molecular network containing m chemical reactions with n molecular species. Let $z = (z_1, ..., z_n)$ be the state of the species, *i.e.*, z_i is the number of the ith species at t, which is a non-negative integer. Define $P(z; t)$

as the probability for the state z at t. Then, the dynamics of the system is described by the master equation with initial state z_0 at $t = 0$, as indicated in (2.56):

$$\frac{\partial P(z;t)}{\partial t} = \sum_{k=1}^{m} [w_k(z - r_k)P(z - r_k;t) - w_k(z)P(z;t)], \qquad (2.224)$$

where $r_k = (r_{k,1}, ..., r_{k,n})$ is an integer vector for the change of state, *i.e.*, $r_{k,j}$ is the change in the number of the jth molecule by the kth reaction. $w_k(z)$ is the transition rate (≥ 0) from state z to state $z + r_k$ by the kth reaction. Notice that $z - r_k$ should be non-negative although $r_{k,j}$ can be negative. Note also that (2.224) is the same as (2.56) although different symbols are used.

In theory, (2.224) provides complete information of system performances, but only a few simple cases are amenable to exact solutions due to its complexity with a large number of variables. Next, we exploit the fast–slow reactions in cellular systems to simplify the master equation.

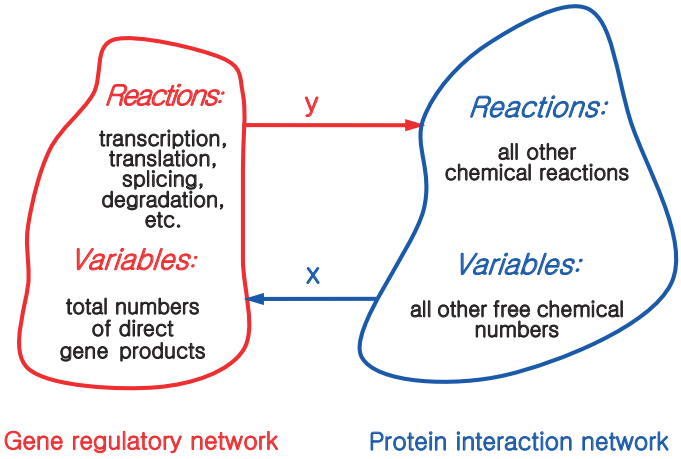

Figure 2.12 Decomposition of a biomolecular network

Although dynamics are intertwined among different biological processes, such as gene regulation and protein interactions, interactions within each biological process are generally more active than those between them. Such a property can be explored to decompose a biomolecular network, in particular, a system, including both a gene regulatory network and a protein interaction network. As shown schematically in Figure 2.12, we define a gene regulatory network (gene network) and a protein interaction network (protein network) in the following manner.

- A gene network is mainly composed of transcription, translation, splicing, and degradation reactions, with total numbers of direct gene products, such as protein and RNA, as variables.

- A protein network is composed of all other chemical reactions, such as phosphorylation, dimerization, binding, enzyme reactions, and chemical modifications, with free chemical numbers as variables.

A gene network involves the gene regulation, and its dynamics are generally slow in contrast to the relatively fast reactions in a protein network. In the gene network, rather than the free monomers, we adopt total numbers of gene products as variables, which include not only free monomers but also those in all complexes, such as dimers and other multimers. Note that the degradation or depletion reactions of direct gene products (chemical monomers) are also included in the gene network. With such definition, the gene–protein network can generally be decomposed into a gene regulatory network and a protein interaction network.

For a general cellular system, there are total m reactions with total n variables. Without loss of generality, assume that the first m_0 reactions are in the protein network. Rearrange the state variables by $z = (x, y)$ and $r_k = (\phi_k, \theta_k)$ with $x = (x_1, ..., x_{n_x})$, $y = (y_1, ..., y_{n_y})$, and $n_x + n_y = n$, where x_i is the number of a molecules synthesized in a fast chemical reaction, and y_i is the total number of mRNA produced by a transcription reaction, or the total number of proteins produced by a translation reaction in the gene network. Notice that y_i for proteins is the total number including those in dimers and other complexes.

Next, we suppress the explicit time dependence of $P(z)$ for readability. Define a marginal function $P(y) = \int_x P(x, y)dx$, where the integration is simply a summation over all discrete x. $P(x)$ is similarly defined as $P(y)$. Then, the joint probability function is written by the marginal probability and the conditional probability as

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x). \tag{2.225}$$

Assume that the concentrations of all house-keeping molecules, such as RNAP and ribosomes, are constant although we can consider them as variables in our model.

We use a simple example to demonstrate the theoretical results. The simple biomolecular network with a single gene is shown in Figure 2.13. A gene is transcribed into mRNAs, which are further translated into protein monomers. Then, the protein monomers are dimerized and act as TFs to regulate the gene activity by binding to the promoter sites. Let numbers of the free protein monomers and the free binding sites of DNA be p and d, respectively. Represent the numbers of the mRNA, the protein dimers, and the protein–DNA complex to be y_1, x_1, and x_2, respectively. Then, the total numbers of the protein and the DNA are $y_2 = p + 2x_1 + 2x_2$ and $u = d + x_2$. Note that using the total numbers as slow variables (*e.g.*, y_2) is more effective than using the number of the free molecules (*e.g.*, p) although it increases the complexity of the formalation.

The protein dimerization and DNA binding reactions are fast dynamics

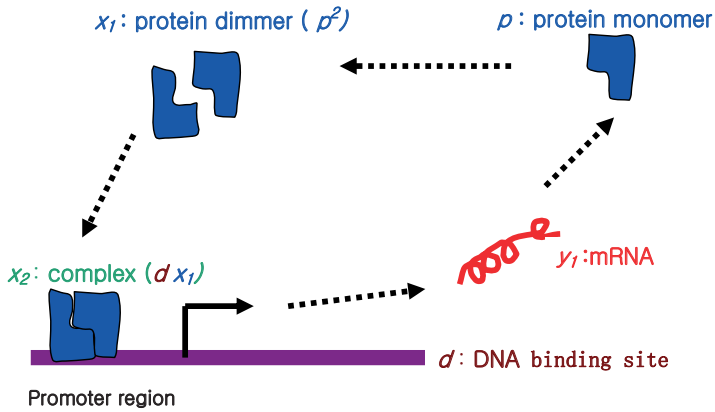

Figure 2.13 A simple molecular network with one gene. Numbers of free protein monomers and free bind sites of DNA are p and d, respectively, whereas total numbers of protein and DNA are $y_2 = p + 2x_1 + 2x_2$ and $d + x_2$, respectively

$$p + p \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} x_1; \quad d + x_1 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} x_2. \tag{2.226}$$

Clearly, there are four fast reactions, including both forward and reverse reactions. On the other hand, the transcription, translation, and degradation reactions are considered as slow dynamics

$$d \overset{\alpha k_3}{\rightharpoonup} y_1 + d; \qquad x_2 \overset{k_3}{\rightharpoonup} y_1 + x_2, \tag{2.227}$$

$$y_1 \overset{k_4}{\rightharpoonup} p + y_1; \qquad y_1 \overset{d_1}{\rightharpoonup} \emptyset; \qquad p \overset{d_2}{\rightharpoonup} \emptyset. \tag{2.228}$$

The transcription reactions occur at different rates for cases of with and without the binding of the dimer x_2 to the promoter, where α can be a repression ($\alpha < 1$) or an activation ($\alpha > 1$) coefficient. There are five slow reactions represented by (2.227)–(2.228).

Let $z = (x, y) = (x_1, x_2, y_1, y_2)$, where $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Therefore, $n = 4$, $m = 9$, $m_0 = 4$, $n_x = 2$, and $n_y = 2$. The terms for $k = 1, ..., 4$ in Table 2.3 correspond to the four reactions in (2.226), whereas the terms for $k = 5, ..., 9$ are derived from the five reactions in (2.227)–(2.228). Then, the master equation (2.224) is obtained directly from Table 2.3 for the simple molecular network.

Clearly, ϕ_k for $k = m_0 + 1, ..., m$ and θ_k for $k = 1, ..., m_0$ are zero vectors. The total numbers of direct gene products (*i.e.*, y_1, and p) are affected only by transcription, translation, and degradation reactions, which are all in the gene network, whereas other chemical numbers vary only in the protein network although there exist interactions between gene and protein networks. Then, the protein network can be described as

Table 2.3 r_k and w_k for the simple molecular network, where V stands for the cell volume and $r_k = (\phi_k, \theta_k) = (\phi_{k,1}, \phi_{k,2}, \theta_{k,1}, \theta_{k,2})$. Similarly, w_1 should be $w_1 = k_1 p(p-1)/V$, rather than $k_1 p^2/V$. $u = d + x_2$ is the total number of binding sites for the gene, which is a constant. θ_k for all $k = 1, ..., 4$ and ϕ_k for all $k = 5, ..., 9$ are zero vectors

k	$r_k = (\phi_k, \theta_k)$	$w_k(x, y)$
1	1,0, 0,0	$k_1 p^2/V = k_1(y_2 - 2x_1 - 2x_2)^2/V$
2	-1,0, 0,0	$k_{-1}x_1$
3	-1,1, 0,0	$k_2 x_1 d/V = k_2 x_1(u - x_2)/V$
4	1,-1, 0,0	$k_{-2}x_2$
5	0,0, 1,0	$\alpha k_3 d = \alpha k_3(u - x_2)$
6	0,0, 1,0	$k_3 x_2$
7	0,0, 0,1	$k_4 y_1$
8	0,0, -1,0	$d_1 y_1$
9	0,0, 0,-1	$d_2 p = d_2(y_2 - 2x_1 - 2x_2)$

$$\frac{\partial P(x|y)}{\partial t} = \sum_{k=1}^{m_0} [w_k(x - \phi_k, y)P(x - \phi_k|y) - w_k(x, y)P(x|y)]. \quad (2.229)$$

Substituting (2.225) into (2.224) and summing over all x, we have the following evolution equation of the marginal functions for the gene network:

$$\frac{\partial P(y)}{\partial t} = \sum_{k=m_0+1}^{m} [\bar{w}_k(y - \theta_k)P(y - \theta_k) - \bar{w}_k(y)P(y)], \quad (2.230)$$

where

$$\bar{w}_k(y) = \sum_x w_k(x, y)P(x|y) \quad (2.231)$$

is an average value conditional to y. $\bar{w}_k(y)$ can be expressed by conditional moments or cumulants of x because $w_k(x, y)$ is generally a polynomial of x and y.

According to (2.225), we can clearly obtain the dynamics of the biomolecular network by (2.229) and (2.230), which is much simpler than the original (2.224) and thus is a reduced biomolecular network. From the viewpoint of computational complexity, the decoupled master equations (2.229)–(2.230) require considerably less CPU demand than the original master equation (2.224) does.

2.7.2 Approximation of Continuous Variables in Molecular Networks

When we approximate x as continuous variables, (2.229) can be expressed approximately by the Fokker–Planck equation

$$\frac{\partial P(x|y)}{\partial t} = -\frac{\partial}{\partial x}\left[A(x)P(x|y)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[B(x)P(x|y)\right], \tag{2.232}$$

where

$$A(x) = (A_1, ..., A_{m_0}), \tag{2.233}$$
$$D(x) = (D_1, ..., D_{m_0}), \tag{2.234}$$
$$B(X) = [B_{ij}]_{m_0 \times n}, \tag{2.235}$$

with

$$A_i(x) = \sum_{j=1}^{n} \phi_{ji} w_j(x, y), \tag{2.236}$$
$$B_{ij}(x) = \phi_{ji} w_j^{1/2}(x, y), \tag{2.237}$$
$$D_i(x) = \sum_{j=1}^{n} B_{ij}^2(x). \tag{2.238}$$

Therefore, we have (2.232) and (2.230) to represent the original (2.224). Generally, compared with the master equation (2.229), $P(x|y)$ can be calculated more efficiently by (2.232) for a given y, thereby yielding $\bar{w}_k$ of (2.231) efficiently. Note that y is still considered to be discrete even when solved by the reduced master equation (2.230). Clearly, (2.229) and (2.232) are a hybrid system with both discrete and continuous variables.

2.7.3 Gaussian Approximation in Molecular Networks

When focusing on the gene network, we can easily examine the dynamics if the conditional moments or cumulants of x are provided, according to (2.230). In other words, provided that the necessary moments or cumulants in $\bar{w}_k(y)$ and $\bar{w}_k(y - \theta_k)$ are available, the dynamics of the gene network simply follow (2.230) with a much smaller number of variables. The required moments or cumulants can be calculated from (2.229) by multiplying x^i for appropriate integer i and summing over all x. Next, we approximate the model (2.224) with (2.230) by assuming that all variables follow approximately Gaussian distribution.

Let N be a conditional mean vector with elements $N_i = \langle x_i|y\rangle$, and M be a conditional correlation matrix with elements $M_{ij} = \langle (x_i - N_i)(x_j - N_j)\rangle$ for x. Then, from (2.231), by expanding w_k at N, the transition rate $\bar{w}_k(y)$ of (2.230) is given as

$$\begin{aligned}\bar{w}_k(y) &= \sum_x [w_k(N, y) + \frac{\partial w_k(N, y)}{\partial x}(x - N) + \cdots]P(x|y) \\ &= w_k(N, y) + \frac{1}{2}\frac{\partial^2 w_k(N, y)}{\partial x^2} M + \cdots \\ &\triangleq h_k(N, M, y),\end{aligned} \tag{2.239}$$

which requires the conditional moments of x. Notice that all odd central moments vanish and the even moments can be expressed by the second central moment M in the Gaussian distribution. Therefore, $\bar{w}_k(y)$ can be expressed by N and M with y, *i.e.*, $h_k(N, M, y)$, and we only need to derive the means and correlations of x when assuming the Gaussian distribution, as indicated in (2.239).

On the other hand, by multiplying x_i by (2.229) and summing over x, we have the following evolution equation for the means:

$$\begin{aligned}\frac{\mathrm{d}N_i}{\mathrm{d}t} &= \sum_{k=1}^{m_0}\sum_{x}\phi_{ki}w_k(x-\phi_k,y)P(x-\phi_k|y)\\ &= \sum_{k=1}^{m_0}\sum_{x}\phi_{ki}w_k(x,y)P(x|y)\\ &= \sum_{k=1}^{m_0}\phi_{ki}\bar{w}_k(y)\\ &\triangleq f_i(N,M,y),\end{aligned} \tag{2.240}$$

where $w_k(x, y)$ is expanded at N in (2.240). According to (2.239), it is obvious that $f_i(N, M, y) = \sum_{k=1}^{m_0}\phi_{ki}h_k(N, M, y)$.

Similarly, we can derive the evolution equations for the correlations by multiplying $(x_i - N_i)(x_j - N_j)$ by (2.229) and summing over x as follows:

$$\begin{aligned}\frac{\mathrm{d}M_{ij}}{\mathrm{d}t} &= \sum_{k=1}^{m_0}\sum_{x}[\phi_{ki}\phi_{kj} + \phi_{ki}(x_j - N_j) + \phi_{kj}(x_i - N_i)]w_k(x,y)P(x|y)\\ &= \sum_{k=1}^{m_0}[\phi_{ki}\phi_{kj}\bar{w}_k(y) + \phi_{ki}\frac{\partial w_k(N,y)}{\partial x}M_j + \phi_{kj}\frac{\partial w_k(N,y)}{\partial x}M_i + \cdots]\\ &\triangleq g_{ij}(N,M,y),\end{aligned} \tag{2.241}$$

where M_i is the ith column of M. Note that M is a symmetrical matrix, *i.e.*, $M_{ij} = M_{ji}$. Therefore, the evolution equations can be expressed in terms of N and M for given y, as indicated in (2.240) and (2.241).

For the simple molecular network shown in Figure 2.13, clearly from (2.239) we can derive the $\bar{w}_1(y)$ shown in Table 2.4, where V is the cellular volume. The evolution equations for means and correlations can easily be derived from (2.240) and (2.241). Therefore, the master equation (2.224) is reduced to a reduced master equation (2.230) with a set of ordinary differential equations (ODEs) (2.240) and (2.241), which can be simulated easily from the numerical computation viewpoint, *e.g.*, using the Gillespie algorithm (Chen *et al.* 2005), due to small numbers of variables and reactions in the reduced master equation.

Therefore, we have a decomposed system that is composed of a master equation (2.230) for the gene network and time-varying ODEs (2.240) and

Table 2.4 Each term for the simple molecular network

$\bar{w}_1(y)$	$k_1(y_2 - 2N_1 - 2N_2)^2/V$
	$+4k_1(2M_{12} + M_{11} + M_{22})/V$
$\bar{w}_2(y)$	$k_{-1}N_1$
$\bar{w}_3(y)$	$k_2N_1(u - N_2)/V - k_2M_{12}/V$
$\bar{w}_4(y)$	$k_{-2}N_2$
$\bar{w}_5(y)$	$\alpha k_3(u - N_2)$
$\bar{w}_6(y)$	k_3N_2
$\bar{w}_7(y)$	k_4y_1
$\bar{w}_8(y)$	d_1y_1
$\bar{w}_9(y)$	$d_2(y_2 - 2N_1 - 2N_2)$
f_1	$\bar{w}_1(y) - \bar{w}_2(y) - \bar{w}_3(y) + \bar{w}_4(y)$
f_2	$\bar{w}_3(y) - \bar{w}_4(y)$
g_{11}	$\bar{w}_1(y) + \bar{w}_2(y) + \bar{w}_3(y) + \bar{w}_4(y)$
	$-8k_1(y_2 - 2N_1 - 2N_2)(M_{11} + M_{12})/V - 2k_1M_{11}$
	$-2k_2(u - N_2)M_{11}/V + 2k_2N_1M_{12}/V + 2k_2M_{12}$
g_{22}	$\bar{w}_3(y) + \bar{w}_4(y)$
	$2k_2(u - N_2)M_{12}/V - 2k_2N_1M_{22}/V - 2k_2M_{22}$
g_{12}	$-\bar{w}_3(y) - \bar{w}_4(y)$
	$-k_2(u - N_2)M_{12}/V + k_2N_1M_{22}/V + k_2M_{22}$
	$k_2(u - N_2)M_{11}/V - k_2N_1M_{12}/V - k_2M_{12}$

(2.241) for the protein network. For an extreme case, assuming that the dynamics of the protein network is much faster than that of the gene network, the moments can be numerically or even analytically calculated by considering (2.240) to be zero. Note that we do not require the quasi-steady-state equilibrium of the fast–slow dynamics in this section for the Gaussian approximation. In other words, with the assumption of the Gaussian distribution, (2.230) with (2.240)–(2.241) is an exact representation of the original master equation (2.224).

2.7.4 Deterministic Approximation in Molecular Networks

We further approximate the model (2.230) with (2.240) and (2.241) by a deterministic scheme in this section. The dynamics of the protein network is relatively fast and the number of variables is usually substantially bigger than that of the gene network. Let V_p be the size of the protein system, *e.g.*, the cellular volume. Since the stochastic deviation is approximately proportional to $1/\sqrt{V_p}$, we assume that the noise of the protein network is almost averaged. In other words, we assume that the reactions in the protein network are deterministic. Therefore, when y is given, according to (2.240), $\langle x|y\rangle = x$ and the protein network for x is a deterministic system, *i.e.*,

$$\dot{x}_i = \sum_{k=1}^{m_0} \phi_{k,i} w_k(x, y), \tag{2.242}$$

where y is given. For y and $x(0)$, we have

$$\dot{\psi}_i(y) = \sum_{k=1}^{m_0} \phi_{k,i} w_k(\psi(y), y) \quad \text{with} \quad \psi(y, 0) = x(0), \tag{2.243}$$

where ψ is the flow of the dynamics. (2.243) is a time-varying ODE due to the time-dependent $y(t)$.

Hence, the conditional probability of x is

$$P(x|y) = \delta(x - \psi(y)) \tag{2.244}$$

and

$$P(x, y) = P(x|y)P(y) = \delta(x - \psi(y))P(y). \tag{2.245}$$

Therefore, we have the following master equation (2.246) with ODE (2.243) which is a hybrid system with both stochastic and deterministic processes:

$$\frac{\partial P(y)}{\partial t} = \sum_{k=m_0+1}^{m} [w_k(\psi(y - \theta_k), y - \theta_k)P(y - \theta_k) - w_k(\psi(y), y)P(y)]. \tag{2.246}$$

For the simple molecular network, the master equation is (2.246) with $m_0 = 4$ and $m = 9$, and ODEs are

$$\frac{\mathrm{d}\psi_2}{\mathrm{d}t} = w_1(\psi, y) - w_2(\psi, y) - w_3(\psi, y) + w_4(\psi, y), \tag{2.247}$$

$$\frac{\mathrm{d}\psi_3}{\mathrm{d}t} = w_3(\psi, y) - w_4(\psi, y), \tag{2.248}$$

where $x_1(t) = \psi_2(t)$ and $x_2(t) = \psi_3(t)$.

Moreover, approximating (2.246) by the Fokker–Planck equation, we equivalently transform (2.246) into the form of the Langevin equations

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = K_i(y) + \eta_i, \tag{2.249}$$

where η_i are Gaussian noise signals with zero mean $\langle \eta_i(t) \rangle = 0$ and covariances $\langle \eta_i(t)\eta_j(t') \rangle = K_{ij}\delta(t - t')$. K_i and K_{ij} are defined by

$$K_i(y) = \sum_{k=m_0+1}^{m} \theta_{k,i} w_k, \tag{2.250}$$

$$K_{ij}(y) = \sum_{k=m_0+1}^{m} \theta_{k,i}\theta_{k,j} w_k, \tag{2.251}$$

for all i and j. When there is no noise, it is easy to check that $\mathrm{d}y_i/\mathrm{d}t = K_i(y)$ is identical to the rate equation of the deterministic system.

Thus, for the simple molecular network, we have the corresponding Langevin equations of (2.249) as follows:

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = K_1(\psi, y) + \eta_1, \tag{2.252}$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = K_2(\psi, y) + \eta_2, \tag{2.253}$$

where η_i are Gaussian noise signals with zero mean $\langle \eta_i(t) \rangle = 0$ and covariances $\langle \eta_i(t)\eta_j(t') \rangle = K_{ij}\delta(t-t')$. $K_1(\psi, y) = w_5 + w_6 - w_8$ and $K_2(\psi, y) = w_7 - w_9$. $K_{11}(\psi, y) = w_5 + w_6 + w_8$ and $K_{22}(\psi, y) = w_7 + w_9$. $K_{12} = K_{21} = 0$. Hence, the gene–protein network can be simplified as (2.247)–(2.248) and (2.252)–(2.253).

As another decomposition method for the general molecular network (2.224), we can also partition the variables z as $z = (x, y)$, where x is the species in large numbers and y represents the species in small numbers. Then, by the partial Taylor expansion or the Kramers–Moyal expansion for x (Crudu *et al.* 2009), we can express the master equation approximately with x as continuous variables but considering y as a discrete variable. In such a way, the computational efficiency can be significantly improved due to a reduced number of discrete variables, which follow the master equation.

2.7.5 Prefactor Approximation of Deterministic Representation

A continuous approximation scheme, which reduces the number of dimensions of a system while predicting the dynamics of the entire system more accurately than that provided by the classic quasi-steady-state (QSS) approximation, was developed (Kepler and Elston 2001, Bundschuh *et al.* 2003, Bennett *et al.* 2007) for the deterministic model. By correctly applying multiple time scale analysis, we found that the resulting reduced systems were similar to the QSS approximations, but with a prefactor in front of the time derivatives of the concentrations. Consider the example of the synthetic repressilator shown in Figure 2.14 (Elowitz and Leibler 2000). The reactions in the repressilator are as follows (Bennett *et al.* 2007):

$$x_i + x_i \underset{\kappa_-}{\overset{\kappa_+}{\rightleftharpoons}} y_i, \tag{2.254}$$

$$d_{i,0} + y_k \underset{k_-}{\overset{k_+}{\rightleftharpoons}} d_{r,i}, \tag{2.255}$$

$$d_{i,0} \overset{\alpha}{\rightharpoonup} d_{i,0} + m_i, \tag{2.256}$$

$$m_i \overset{\sigma}{\rightharpoonup} m_i + x_i, \tag{2.257}$$

$$m_i \overset{\gamma_{\mathrm{m}}}{\rightharpoonup} 0, \tag{2.258}$$

$$x_i \overset{\gamma_{\mathrm{p}}}{\rightharpoonup} 0, \tag{2.259}$$

where $d_{i,0}$ and $d_{r,i}$ are the free and repressed promoter sites, respectively. If the promoter of gene G_i is free, it can transcribe its associated mRNA, m_i, which in turn can translate its associated protein.

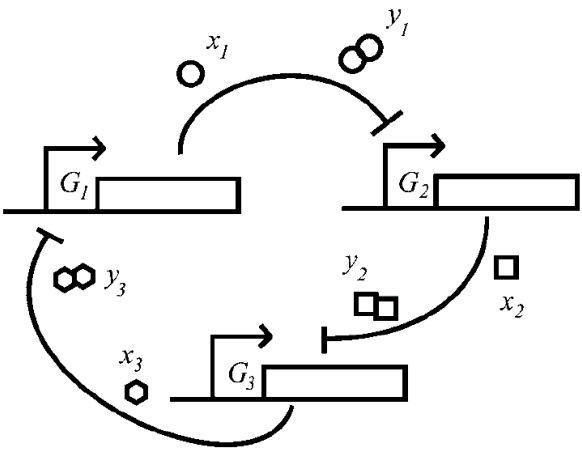

Figure 2.14 Schematic representation of the repressilator. Gene G_1 produces protein x_1, whose dimer y_1 inhibits transcription of gene G_2. Similarly, the protein dimer y_2 represses the gene G_3, whose protein dimer y_3 represses transcription of G_1 (from (Bennett *et al.* 2007))

According to the mass action law, the differential equations are given by

$$\dot{x}_i = -2\kappa_+ x_i^2 + 2\kappa_- y_i + \sigma m_i - \gamma_p x_i, \tag{2.260}$$

$$\dot{y}_i = \kappa_+ x_i^2 - \kappa_- y_i - k_+ y_i d_{0,j} + k_- d_{r,j}, \tag{2.261}$$

$$\dot{d}_{i,0} = -k_+ d_{i,0} y_k + k_- d_{r,i}, \tag{2.262}$$

$$\dot{d}_{r,i} = k_+ d_{i,0} y_k - k_- d_{r,i}, \tag{2.263}$$

$$\dot{m}_i = \alpha d_{i,0} - \gamma_m m_i, \tag{2.264}$$

where $i \in \{1,2,3\}$, $j \in \{2,3,1\}$, and $k \in \{3,1,2\}$.

Assume that the dimerization and dissociation processes are faster than the other processes and that these reactions are in equilibrium. Solving the resulting algebraic equations, we obtain

$$y_i = c_p x_i^2, \tag{2.265}$$

$$d_{i,0} = \frac{d}{1 + c_d c_p x_k^2}, \tag{2.266}$$

$$d_{r,i} = \frac{d c_d c_p x_k^2}{1 + c_d c_p x_k^2}, \tag{2.267}$$

where $d = d_{0,i} + d_{r_i}$, $c_p = \kappa_+/\kappa_-$, and $c_d = k_+/k_-$. Substituting these expressions into (2.260)–(2.261), we obtain the following differential equations:

$$\dot{n}_i = \dot{x}_i \frac{\partial n_i}{\partial x_i} = \dot{x}_i p(x_i), \tag{2.268}$$

$$\dot{m}_i = \frac{\alpha d}{1 + c_d c_p x_k^2} - \gamma_m m_i, \tag{2.269}$$

where $n_i = x_i + 2y_i + 2d_{r,j} \approx x_i + 2c_p x_i^2 + 2c_d c_p d x_i^2 (1 + c_d c_p x_i^2)^{-1}$ and

$$p(x_i) = 1 + 4c_p x_i + \frac{4c_d c_p d x_i}{(1 + c_d c_p x_i^2)^2}. \tag{2.270}$$

Clearly n_i is the total number or concentration of protein i, including those in the complexes. Defining the dimensionless variables by rescaling $\gamma_m t \to t$, $\sqrt{c_d c_p x_i} \to x_i$, and $(\sigma\sqrt{c_d c_p})/(\gamma_m \beta) m_i \to m_i$, we obtain the system

$$p(x_i)\dot{x}_i = \beta(m_i - x_i), \tag{2.271}$$

$$\dot{m}_i = \frac{\kappa d'}{1 + x_k^2} - m_i, \tag{2.272}$$

where $\beta = \gamma_p/\gamma_m$, $\kappa = \alpha\sigma/\gamma_m\gamma_p$, $d' = \sqrt{c_d c_p} d$, and $p(x_i)\dot{x}_i = \dot{n}_i$. Note that $f(x_i)$ in (2.271) is the rescaled one, *i.e.*,

$$p(x_i) = 1 + 4r x_i + \frac{4d' x_i}{(1 + x_i^2)^2}, \tag{2.273}$$

where $r = \sqrt{c_p/c_d}$.

The difference between (2.271)–(2.272) and the model under QSS as used in (Elowitz and Leibler 2000) is the existence of the prefactor $p(x_i)$. Such a difference results from different reduction approaches, *i.e.*, treating x_i and $n_i = x_i + 2y_i + 2d_{r,j}$ as slow variables under the QSS and the prefactor approximation, respectively. It is true that x_i depends on slow reactions, but it also depends on fast reactions. Therefore, x_i is not a slow variable, but a mixture of both slow and fast variables. Here, the true slow variable is n_i, representing the total concentration of protein molecules.

It has been shown that the prefactor approximation predicts the dynamics of the entire system more accurately than that provided by the classic QSS approximation, including both transient dynamics and other properties such as relaxation times of equilibria and periods and amplitudes of oscillations. For example, a comparison of the periods and amplitudes predicted by the entire system (2.260)–(2.264), the prefactor approximation (2.271)–(2.272), and the QSS approximation (*i.e.*, $p(x_i) = 1$ in (2.271)–(2.272)) is shown in Figure 2.15. It can be seen that the prefactor approximation predicts the dynamics of the entire system more accurately than the QSS approximation with respect to the estimation of both the periods and amplitudes. See (Bennett *et al.* 2007) for more theoretical analysis and examples.

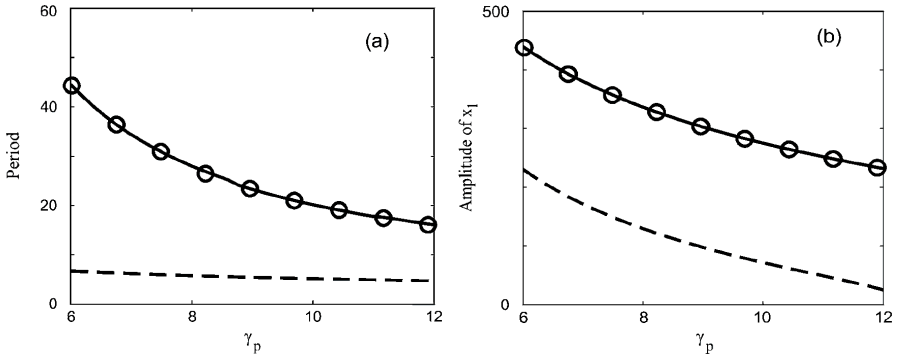

Figure 2.15 A comparison of the periods (a) and the amplitudes (b) of the oscillations as functions of γ_p for the entire system (black curves), the prefactor approximation (circles), and the QSS approximation (dashed curves). The parameters are $\gamma_p = 6$, $\gamma_m = 1$, $\kappa_+ = k_+ = 5$, $\kappa_- = k_- = 100$, $\alpha = 10$, $\sigma = 20$, and $d = 20$ (from (Bennett *et al.* 2007))

2.7.6 Stochastic Simulation of Hybrid Systems

Define (x, y) and (X, Y) to be concentrations and numbers of molecules, respectively. Reconsider the general system (2.224) by considering $z = (x, y)$ and $r_k = (\phi_k, \theta_k)$ with $x = (x_1, ..., x_{n_x})$, $y = (y_1, ..., y_{n_y})$, and $n_x + n_y = n$, where x_i and y_i are the concentrations of molecules, *i.e.*, the numbers divided by the system size or volume V. Then, the dynamics of the system is described by the master equation

$$\frac{\partial P(x, y; t)}{\partial t} = \sum_{k=1}^{m} [w_k(x - \phi_k/V, y - \theta_k/V)P(x - \phi_k/V, y - \theta_k/V; t) - w_k(x, y)P(x, y; t)], \tag{2.274}$$

where (ϕ_k, θ_k) is a vector for the change of the state, *i.e.*, $r_{k,j}$ is the change in the number of the jth molecule by the kth reaction. $w_k(x, y)$ is the transition rate (≥ 0) from state (x, y) to state $(x + \phi_k/V, y + \theta_k/V)$ by the kth reaction. Note that (x, y) represent the concentrations in (2.274), in contrast to the numbers z in (2.224).

Hybrid System with Deterministic Process

Assuming that the number of X is much bigger than that of Y, we can approximate X by continuous variables x, *i.e.*, $x = X/V$, by keeping $y = Y/V$ as discrete variables. Therefore, by the partial Kramers–Moyal expansion of (2.274) with respect to x and ϕ_k/V up to the first order (*i.e.*, the zeroth order and the first order), we have the following hybrid representation

$$\frac{\partial P(x,y;t)}{\partial t} = \sum_{k=1}^{m}[w_k(x,y-\theta_k/V)P(x,y-\theta_k/V;t) - w_k(x,y)P(x,y;t)] - \sum_{j=1}^{n_x}\frac{\partial}{\partial x_j}\left[\left(\sum_{k=1}^{m}\phi_{kj}W_k(x,y)\right)P(x,y;t)\right] + \mathcal{O}(\frac{1}{V}), \quad (2.275)$$

where $w_k = W_k V$ is rates of the reactions which are proportional to the volume V based on the mass action law. $\mathcal{O}(\frac{1}{V})$ implies that the order of the term is equal to or higher than $1/V$. Clearly, the term of the Taylor zeroth order expansion, *i.e.*, the first term $\sum_{k=1}^{m}[\cdot]$ in (2.275), is the discrete dynamics or the master equation for discrete variables y, and the term of the Taylor first-order expansion, *i.e.*, the second term $\sum_{j=1}^{n_x}\frac{\partial}{\partial x_j}[\cdot]$ in (2.275), is the deterministic kinetic dynamics or corresponds to the Langevin equation for the continuous variables x. The term of the Taylor second-order expansion, *i.e.*, $\mathcal{O}(\frac{1}{V})$, corresponds to the diffusion process, and when $V \to \infty$, it approaches zero. Therefore, (2.275) is a hybrid system with both discrete and continuous dynamics, or with both stochastic and deterministic processes. Specifically, the stochastic system for discrete variables Y is

$$\frac{\partial P(x,y;t)}{\partial t} = \sum_{k=1}^{m}[w_k(x,y-\theta_k/V)P(x,y-\theta_k/V;t) - w_k(x,y)P(x,y;t)], \quad (2.276)$$

and the deterministic system with continuous variables x is for $j = 1, ..., n_x$

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = \sum_{k=1}^{m}\phi_{kj}W_k(x(t),y), \quad (2.277)$$

where $j = 1, ..., n$. Let $\delta(\theta_k) = 0$ if θ_k is a zero vector; otherwise, $\delta(\theta_k) = 1$. Therefore, defining the jump intensity $w_0 = \sum_{k=1}^{m} w_k(x,y)\delta(\theta_k)$, we have the algorithm of stochastic simulation based on the piecewise deterministic Markov process (PDMP) (Davis 1993, Zeiser *et al.* 2008, Crudu *et al.* 2009), shown in Table 2.5.

Clearly, the continuous variables x are governed by the deterministic system and change continuously during each time interval $[t_i, t_i + \Delta t_i)$, while the discrete variables y remain constant. Therefore, x are piecewise continuous variables and may change discretely at $t_i + \Delta t_i$. On the other hand, y evolve discretely with stochastic motion punctuated by a sequence of random waiting times Δt_i due to the master equation (2.276). Such dynamics is schematically shown in Figure 2.16.

Hybrid System with Diffusion Process

Moreover, we can expand (2.274) to the second order of V to consider the diffusion effect, *i.e.*,

Table 2.5 Algorithm of stochastic simulation for (2.275) based on PDMP

Step 1. Initialization: set $t_0 = 0$ and fix the initial numbers of molecules $(X_0/V, Y_0/V)$.

Step 2. Calculate the propensity function w_k, $k = 1, ..., m$.

Step 3. Generate one random number r_1 uniformly distributed in $[0, 1)$.

Step 4. Integrate the following differential equations:

$$\begin{aligned}\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} &= \sum_{k=1}^{m} \phi_{kj} W_k(x(t), y) \text{ for } j = 1, ..., n_x \\ \frac{\mathrm{d}q(t)}{\mathrm{d}t} &= -w_0(x(t), y)q(t) \\ &\quad \text{with } \ x(t_i) = x_i, \ q(t_i) = 1\end{aligned}$$

between t_i and $t_i + \Delta t_i$ with the stopping condition $q(t_i + \Delta t_i) = r_1$. Then, we have Δt_i.

Step 5. Generate a second random number r_2 which is uniformly distributed in $[0, 1)$. Choose μ_i so that μ_i is the smallest integer satisfying $\sum_{j'=1}^{\mu_i} w_{j'}(x, y) > r_2 w_0(x, y)$.

Step 6. Execute the reaction μ_i, *i.e.*, update (x, y). If $t_i > T_{max}$, terminate the computation. Otherwise, go to Step 2.

$$\frac{\partial P(x, y; t)}{\partial t} = \sum_{k=1}^{m} [w_k(x, y - \theta_k/V)P(x, y - \theta_k/V; t) \tag{2.278}$$

$$-w_k(x, y)P(x, y; t)] \tag{2.279}$$

$$-\sum_{k=1}^{m}\sum_{j=1}^{n_x} \frac{\partial}{\partial x_j} [g_{jk}(x, y; t)P(x, y; t)] \tag{2.280}$$

$$+\sum_{j=1}^{n_x}\sum_{l=1}^{n_n} \frac{\partial^2}{\partial x_j \partial x_k} \left(\sum_{k=1}^{m} \frac{\phi_{kj}\phi_{kl}}{2V} W_k(x, y)P(x, y; t) \right) \tag{2.281}$$

$$+\mathcal{O}(\frac{1}{V^2}) \tag{2.282}$$

where

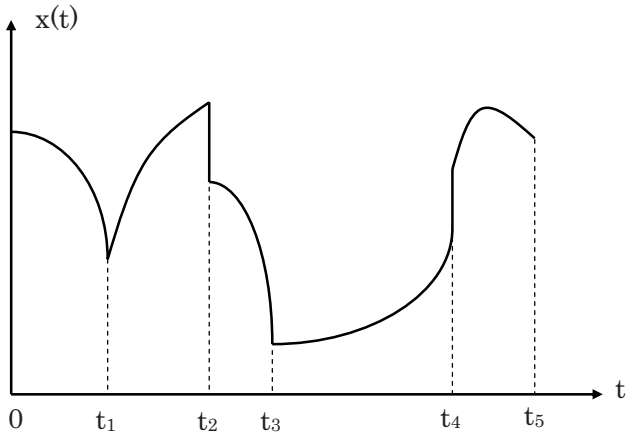

Figure 2.16 Schematic illustration of a piecewise deterministic Markov process

$$g_{jk}(x,y;t) = \phi_{kj}W_k(x,y) - \sum_{l=1}^{n_y} \frac{\phi_{kj}\theta_{kl}}{VP(x,y;t)} \frac{\partial W_k(x,y)P(x,y;t)}{\partial y_l}$$
$$= \phi_{kj}W_k(x,y) - \sum_{l=1}^{n_y} \frac{\phi_{kj}\theta_{kl}}{V} [\frac{\partial W_k(x,y)}{\partial y_l}$$
$$+W_k(x,y)\frac{\partial lnP(x,y;t)}{\partial y_l}]. \quad (2.283)$$

Therefore, (2.280)–(2.281) can also be expressed by the Langevin equations instead of the differential equations (2.277), *i.e.*, stochastic differential equations with the continuous variables x for $j = 1, ..., n$

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = \sum_{k=1}^{m} g_{jk}(x(t),y;t) + \sum_{k=1}^{m} \frac{\phi_{kj}}{\sqrt{V}} \sqrt{W_k(x(t),y)}\Gamma_k(t), \quad (2.284)$$

where $\Gamma_k(t)$ is defined in (2.105). For this case, the hybrid system is the combination of the discrete stochastic system (2.276) and continuous stochastic system (2.284), which can be simulated similarly to the algorithm shown in Table 2.5 and Table 2.6.

In Table 2.6, $V_j(t)$ are independent one-dimensional Wiener processes. The stochastic differential equation can be calculated by the Itô integration. Clearly, there is $P(x(t),y;t)$ in (2.284) or g_{jk}, which is required to be estimated during the integration. There are many ways to approximate $P(x(t),y;t)$, such as by the finite state projection approach, the Gaussian distribution assumption for the continuous variables, or the equilibrium probability distribution.

We consider the scheme to estimate approximately $\partial \ln P(x,y;t)/\partial y_l$. Since y corresponds to discrete variables which are expected to change the

Table 2.6 Algorithm of stochastic simulation for (2.278)–(2.282) based on PDMP

Step 1. Initialization: set $t_0 = 0$ and fix the initial numbers of molecules $(X_0/V, Y_0/V)$.

Step 2. Calculate the propensity function w_k, $k = 1, ..., m$.

Step 3. Generate one random number r_1 uniformly distributed in $[0, 1)$.

Step 4. Integrate the following stochastic differential equations:

$$\begin{aligned} \mathrm{d}x_j(t) &= \sum_{k=1}^{m} g_{jk}(x(t), y; t)\mathrm{d}t + \sum_{k=1}^{m} \frac{\phi_{kj}}{\sqrt{V}} \sqrt{W_k(x(t), y)}\mathrm{d}V_k(t) \\ &\quad \text{for } j = 1, ..., n_x \\ \mathrm{d}q(t) &= -w_0(x(t), y)q(t)\mathrm{d}t \\ &\quad \text{with } \ x(t_i) = x_i, q(t_i) = 1 \end{aligned}$$

between t_i and $t_i + \Delta t_i$ with the stopping condition $q(t_i + \Delta t_i) = r_1$. Then, we have Δt_i.

Step 5. Generate a second random number r_2 which is uniformly distributed in $[0, 1)$. Choose μ_i so that μ_i is the smallest integer satisfying $\sum_{j'=1}^{\mu_i} w_{j'}(x, y) > r_2 w_0(x, y)$.

Step 6. Execute the reaction μ_i, *i.e.*, update (x, y). If $t_i > T_{max}$, terminate the computation. Otherwise, go to Step 2.

dynamics in a slow manner in contrast to the continuous variables x, we assume $\partial P(x(t), y; t)/\partial y_l \approx 0$, or $\partial \ln P(x(t), y; t)/\partial y_l \approx 0$. Specifically, we have

$$g_{jk}(x, y; t) = \phi_{kj} W_k(x, y) - \sum_{l=1}^{n_y} \frac{\phi_{kj}\theta_{kl}}{V} \frac{\partial W_k(x, y)}{\partial y_l}. \tag{2.285}$$

2.8 Stochastic versus Deterministic Representation

Stochastic and deterministic approaches have both advantages and disadvantages. The advantage of stochastic approaches is that they exactly capture the discrete and stochastic nature of cellular systems. At low concentrations of molecules, molecular fluctuations are likely to have a marked impact on system dynamics. The predictions of deterministic and stochastic models for circadian rhythms show that robust circadian oscillations can be observed even when the maximum number of mRNA and protein molecules is of the order of some tens and hundreds (Gonze *et al.* 2002b). To assess exactly the effects of molecular noise, it is necessary to resort to a stochastic approach.

However, almost all analytical methods available for deterministic approaches are no longer applicable and all stochastic simulations are time-consuming processes. The computational efficiency rapidly degrades as the complexity of a system increases. One should therefore use stochastic approaches only if they are absolutely necessary.

Deterministic approaches neglecting the discrete and stochastic nature have received considerable attention due to the simplicity for performing qualitative and quantitative studies. For deterministic formalisms, there are rich techniques, *e.g.*, structural analysis, cellular control analysis, frequency analysis, and bifurcation analysis, that can be used to qualitatively or quantitatively analyze the system dynamics. In many cases, stochastic and deterministic descriptions of a system coincide in the sense that the mean behavior of the system can be accurately captured by the deterministic description. For example, sustained oscillation corresponding to a limit cycle in a deterministic circadian rhythm model can also be obtained in its corresponding stochastic description, *i.e.*, the stochastic oscillations present fluctuations around the deterministic limit cycle (Gonze *et al.* 2006).

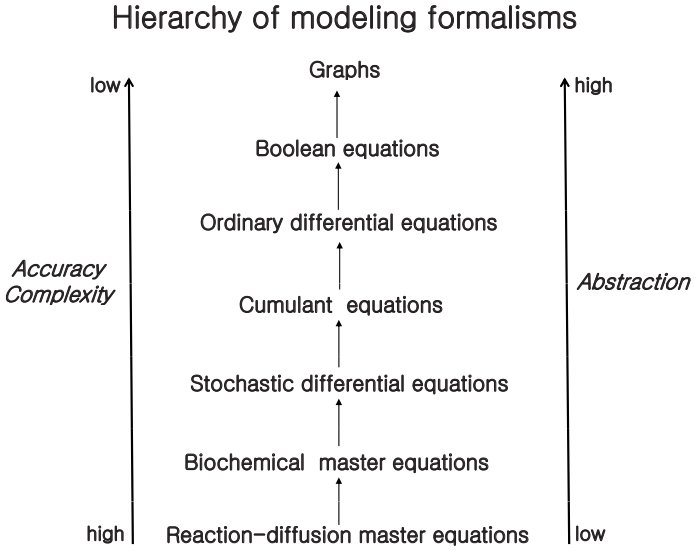

Figure 2.17 Formalisms to model molecular networks

Generally, when a deterministic system is operating near a critical point, stochastic and deterministic processes may be substantially different. In this case, noise can induce some new phenomena or qualitative changes. For example, for some parameters, both the stochastic and deterministic processes in a circadian rhythm model coincide, *i.e.*, they both exhibit periodic oscillations

with only difference in their periods and fluctuations in terms of concentrations. However, for other parameters, the deterministic description results in a stable equilibrium, while stable oscillations persist in the stochastic description (Vilar *et al.* 2002).

Some other noise-induced new phenomena such as noise-based switches and amplifiers for gene expression (Hasty *et al.* 2000) and fluctuation-enhanced sensitivity of intracellular regulation (Paulsson *et al.* 2000) in a single cell have also been developed. Therefore, when species are present at low copy numbers, the stochastic description is more reasonable although it is solvable neither analytically nor with high computational efficiency. On the other hand, when the species numbers are high and the system is operating far from its critical points, the deterministic description is more reasonable due to its simple representation and high computational efficiency. The features of various formalisms to model molecular networks are briefly summarized in Figure 2.17. Depending on the requirement for accuracy, we can choose different modeling approaches for quantitative simulation and qualitative analysis of cellular dynamics.

http://www.springer.com/978-1-84996-213-1

Modeling Biomolecular Networks in Cells
Structures and Dynamics
Chen, L.; Wang, R.; Li, C.; Aihara, K.
2010, XII, 343 p., Hardcover
ISBN: 978-1-84996-213-1