CHAPTER 23

Stochastic Modeling Methods in Cell Biology

Sean X. Sun,[*] Ganhui Lan,[*] and Erdinc Atilgan[†]

[*]Department of Mechanical Engineering and Biomedical Engineering
Johns Hopkins University
Baltimore, Maryland 21218

[†]Department of Anatomy and Structural Biology
Albert Einstein College of Medicine and Yeshiva University
Bronx, New York 10461

Abstract

Stochastic methods have been a staple for understanding complex systems in chemistry and physics. In the biological context, they are useful for understanding phenomena ranging from molecular-level fluctuations to cellular movement. We review the basic formalism behind stochastic methods and outline how they can be implemented for quantifying gene expression, movement of molecular motors, and the dynamics of cytoplasmic components. We show that stochastic methods are quantitative checks for proposed molecular mechanisms and can pose new

METHODS IN CELL BIOLOGY, VOL. 89

0091-679X/08 $35.00
DOI: 10.1016/S0091-679X(08)00623-7

questions for experiments. Structural information of cellular components can be incorporated into stochastic models to reveal new biological insights.

I. Introduction

As the name suggests, stochastic methods have an element of randomness that gives rise to probabilistic behavior. However, the justification of the method is firmly rooted in physics, and is an emerging principle in systems with many reacting/interacting components. The method is equally applicable to analyzing conformational transitions in a single molecule, or reaction dynamics in a cell with millions of proteins. The general argument starts from the microscopic equation of motion for atoms and molecules. Let us first consider a single protein in solution (Fig. 1). Typically, we are not interested in the motions of all the components (e.g., solvents), but are only interested in the motions of a selected set of degrees of freedom (e.g., the angle between two domains in Fig. 1). From the deterministic Newton's equation for all the atoms in the system, one can derive equations of motion for the select set (Hanggi *et al.*, 1990):

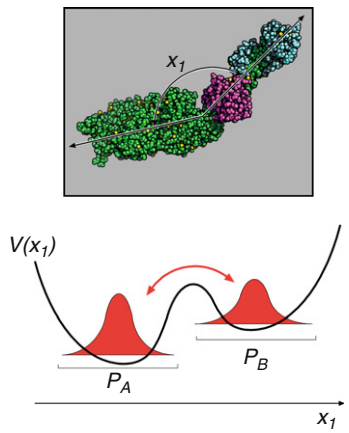

Fig. 1 In a complex system, only a few degrees of freedom are of specific interest. For example, this protein in solution undergoes a conformational change and the value of x_1 is the quantity of interest. The effective equation of motion for x_1 is given in Eq. (1). The free energy of the system as a function of x_1 typically shows several local minima (lower panel). The system fluctuates between these energetic basins (reactants and products), and these fluctuations are described by time dependent probability distributions (red distributions). Instead of considering continuous probability distributions, we can consider discrete probabilities, P_A and P_B, which are the integrals of the distributions over the basin regions. The dynamics of P_A and P_B are described by the Markov model in Eq. (2). Stochastic simulations of the model will give trajectories that jump between the basins.

$$\gamma \frac{\partial x_i}{\partial t} = -\frac{\partial V(x_1, x_2, \ldots)}{\partial x_i} + R(t) \quad (1)$$

where $V(x_1, x_2, \ldots)$ is the free energy as a function of the selected degrees of freedom while the other uninteresting degrees of freedom are in thermal equilibrium; γ is a friction constant related to energy loss to the uninteresting degrees of freedom; $R(t)$ is a random force representing the random bombardment of solvent molecules on x_i. When $V(x_1, x_2, \ldots)$ is zero or a constant, the familiar equation for the Brownian motion of a structureless particle results. However, $(x_1, x_2, \ldots)$ can be much more general. For example, they could represent the conformations of a protein or the coordinates of a substrate in an enzyme. The qualitative features of $V(x_1, x_2, \ldots)$ typically show several metastable minima, separated by barriers (Fig. 1). If there are many interesting degrees of freedom, then $V(x_1, x_2, \ldots)$ is a function of many variables too complex to be drawn here.

For typical biomolecular applications inside the cell, friction and random forces are dominant factors; trajectories from Eq. (1) show large fluctuations. The probability of observing a particular trajectory depends on the random force history. Therefore, it is more convenient to consider a time-dependent probability distribution, $P(x_1, x_2, \ldots, t)$, representing many trajectories averaged over the random forces. The evolution of $P(x_1, x_2, \ldots, t)$ on the energy landscape $V(x_1, x_2, \ldots)$ is described by the Fokker–Planck equation, which can be derived from Eq. (1) from first principles (Risken, 1989). Figure 1 shows a hypothetical energy landscape with two minima. If $P(x_1, x_2, \ldots, t)$ at $t = 0$ is concentrated at one of the minima, one will see an exponential increase of the probability in the other minima. The exponential time constant is related to a "rate constant" familiar from kinetic theories. A molecular theory of reaction rate can be derived from statistical mechanics and the fluctuation–dissipation theorem, and is discussed elsewhere (Chandler, 1987).

Behaviors of probability distributions separated by barriers are relatively independent of the detailed features of the energy landscape. The forward and reverse rate constants are most sensitive to the height of the barrier and the features of the barrier immediately surrounding the transition state (reaction bottleneck) (Hanggi *et al.*, 1990). If we integrate over regions of the probability distribution near the basins of stable states and re-label these regions as P_A, P_B, etc, where A and B now label these energetic basins, then the time-dependent behavior of $P_A(t)$ and $P_B(t)$ can be described by a generic set of dynamic equations, the so-called Master equation,

$$\frac{dP_A}{dt} = -k_1 P_A + k_2 P_B \quad \frac{dP_B}{dt} = k_1 P_A - k_2 P_B \quad (2)$$

These basins described by A and B are now recognized as chemical species, or reactants and products. Fluxes between these basins are related to forward and reverse rate constants k_1 and k_2, respectively. Note that this development simplifies the discussion of probability flow on the energy landscape; instead of considering a

continuous probability distribution, we are reduced to discrete probabilities of two states. The trade-off is the introduction of parameters k_1 and k_2, which typically must be measured from experiments; only very rarely can one compute the rate constants from first principles. Equation (2) is a simple 2-state *Markov chain* model and describes a *single* conversion process where one reactant is converted into one product. Analytic solutions to Eq. (2) are readily available.

The forward and reverse rate constants, k_1 and k_2, are related to the flux across the barrier. At steady-state, forward and reverse fluxes must be balanced to obtain a steady equilibrium probability distribution. The equilibrium probability distribution, in turn, depends on the relative energy difference between A and B. Thus, the condition of detailed balance requires that the ratio of the rate constants must be, $k_1/k_2 = e^{-(F_B - F_A)/k_B T}$, where $F_{A,B}$ are the free energies of A and B, respectively. k_B is the Boltzmann constant and T is the temperature. (At room temperature, $k_B T$ is roughly 0.6 kcal/mol.) This requirement ensures that Eq. (2) reproduces the correct equilibrium distribution at steady state. Suppose a force is applied along the x_1 direction in Fig. 1, then the energy landscape would be changed, and the relative energies of A and B would be changed. The forward and reverse rate constants then must be changed to reflect the energy change. A simple model to account for this is the Bell model (Bell, 1978; Hummer and Szabo, 2003). This issue is important whenever forces are present in biomolecular systems.

In more complex situations, there may be more than two basins in the energy landscape. Probabilities in these basins, P_A, P_B, P_C, etc., can flow among each other, resulting in a multi-state Markov model. Such models have been used to describe motions of molecular motors seen in single molecule experiments. Therefore, a multi-state Markov model is usually the basic starting point for stochastic simulation methods. Let us consider a simple multistate example. Let us take the example in Eq. (2), but for N reacting molecules shown in Fig. 2. The Master equation approach is still applicable, but we must redefine the definition of a Markov state: $s = (N_A, N_B)$ where N_A and N_B are the number of reactants and products, respectively. For example, if there are 10 molecules, then $N_A + N_B = 10$. The total number of states is 11, e.g., $s = (10,0)$ is a state, as well as $s = (9,1)$ and $s = (1,9)$, etc. Instead of 2 equations, we have 11 equations, one for each P_s. The rate constants k_1 and k_2 again describe conversions between states, e.g., $s = (10,0)$ can become $s = (9,1)$ with rate k_1, and $s = (9,1)$ can become $s = (10,0)$ with rate k_2. Not all of the states can directly convert to others; for example, $s = (5,5)$ cannot become $s = (2,8)$ with one reaction. However, $s = (5,5)$ can be converted to $s = (2,8)$ with three successive reactions, or jumps. Figure 2 shows one possible sequence of reactions in this system. The probability of observing the system in a particular state is obtained by solving a system of equations such as Eq. (2). Since there are now 11 states, Eq. (2) becomes a set of equations

$$\frac{dP_s}{dt} = -(k_1 + k_2)P_s + k_1 P_{s-1} + k_2 P_{s+1} \tag{3}$$

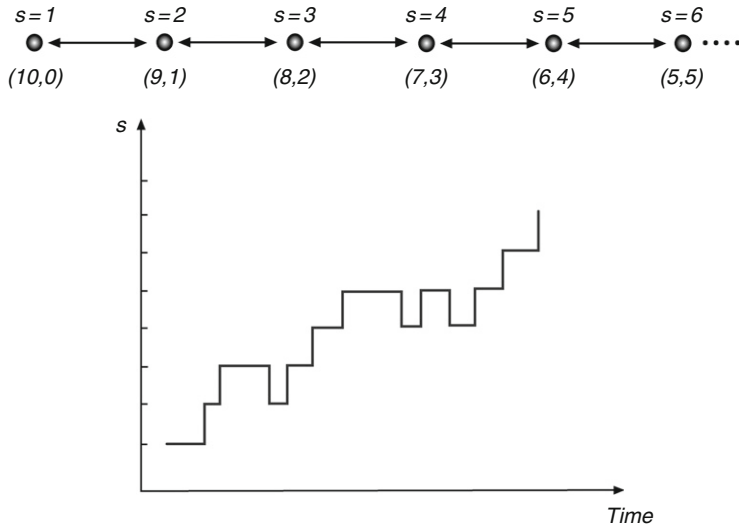

Fig. 2 A multistate Markov model where each state is characterized by the number of reactants and products, N_A and N_B, respectively. For example, $s = 1$ labels $N_A = 10$ and $N_B = 0$. The probability to be in a particular state is described by Eq. (3). A stochastic simulation will show jumps between states and stochastic changes in N_A and N_B.

where s denotes the index of the state. Note that this equation needs to be modified slightly for $s = 1$ and $s = 11$, since there are no $s - 1$ and $s + 1$ states, respectively. Solutions of Eq. (3) give the probability of observing a particular combination of reactants and products. It can be written in a matrix form and implemented using ordinary differential equation solvers in standard packages such as MATLAB.

So far, we have discussed a relatively general model of probability flow between energetic basins using the concept of Markov states. Indeed, the concept is not limited to reacting species because the definition of a Markov state can be much broader. A state can be generalized to regions of the energy surface even when there are no barriers separating them. For example, a diffusing particle can be modeled by a diffusion equation or by a Markov model where the particle hops on a lattice (Wang *et al.*, 2003). The rate of hopping is related to the diffusion constant. In this way, it is possible to model reactions and diffusions of a set of proteins and examine their spatial and temporal dynamics. Spatial hopping algorithms together with reactions have been applied to the MinCDE system in *Escherichia coli* to quantify the spatial–temporal oscillations of these proteins (Kerr *et al.*, 2006; Lutkenhaus, 2007).

When the number of reacting species is large, the number of Markov states increases. This leads to many equations of the form of Eq. (3), and can become prohibitively difficult to solve numerically. Therefore, instead of considering

reactions of individual molecules, we can examine the changes in species concentration. In the simple case of Eq. (2), the concentrations of A and B, C_A and C_B, are related to the probabilities P_A and P_B by a unit conversion. By considering the number of collisions between the reacting partners, familiar kinetic equations emerge. Therefore, the equations for C_A and C_B are the same as Eq. (2) and are the familiar kinetic equations for conversion between A and B. For more complicated situations involving many reacting species, kinetic equations can become nonlinear and exhibit complex and interesting dynamics. Kinetic models can be understood as a mean field approximation to the Master equation where instead of examining microscopic Markov states an equation for a set of averaged quantities is computed. This chapter will not discuss kinetic models, other than to point out that Markov Master equations models can be considered a detailed version of kinetic models. The reader can refer to many available textbooks on kinetic models (e.g., Murray, 1989; Reichl, 1998).

Solutions to equations such as Eq. (3) give the probability of observing a particular Markov state in time, $P_s(t)$. In an experiment, a sequence of states is usually measured. Only after averaging over many measurements, or an ensemble of the same event, does one obtain the probability of observing a particular state. Can we directly compare a single experiment with Eq. (3)? The answer is to use stochastic simulations. A method that is consistent with Eq. (3) was developed in the context of magnetic spin dynamics (Bortz *et al.*, 1975). The same stochastic simulation method applied to chemical reactions was developed by Gillespie (1977), and in this chapter, we shall refer to the algorithm as the Gillespie algorithm. We note that the Gillespie algorithm is a form of importance sampling, much like the Metropolis Monte Carlo algorithm in equilibrium simulations. The algorithm generates probable sequences of states and jump times (trajectories). The probability distribution sampled by the Gillespie algorithm is a time-dependent nonequilibrium distribution consistent with equations such as Eq. (3) (Sun, 2006). The procedure for generating Gillespie trajectories is described in Section V.

Since the algorithm generates stochastic trajectories according to their probability, we can compute observables by directly averaging over the trajectories. Suppose we are interested in measuring a particular quantity, $Z(t)$, which could be a function of time, we can directly evaluate:

$$Z(t) = \frac{1}{N}\sum_{i=1}^{N} Z_i(t) \tag{4}$$

where N is the number of trajectories and $Z_i(t)$ is the quantity of interest evaluated for the ith trajectory. The average would be more accurate if N is large. It is important to note that the algorithm is efficient when computing averages such as Eq. (4). However, just as in Metropolis Monte Carlo, the algorithm does not give the absolute probability of states, and therefore cannot compute $P_s(t)$ in Eq. (3). Nevertheless, we are often interested in average quantities and the Gillespie

algorithm is simple to implement. In the following sections, we discuss stochastic simulations based on the Gillespie algorithm in the context of several cell biological situations.

For some applications, Eq. (1) can be solved directly if the free energy surface is known. Algorithms for obtaining the trajectories can be found in textbooks (Frenkel and Smit, 1996). This type of simulation is often used in molecular dynamics and will not be discussed extensively here.

II. Stochastic Methods in Signaling and Genetic Networks

With improving experimental assays, it is possible to quantify many aspects of protein expression and signaling events in eukaryotic and prokaryotic cells. The number of copies of a particular gene, or a particular phosphorylated kinase can be quite small in a cell. The measured transcriptional activity sometimes shows quantitative differences from cell to cell. Variations in the transcriptional levels may be understood in terms of the stochastic model outlined above. Stochastic effects are especially relevant in bacteria where the regulatory control of gene systems is not rigorous. These issues have been discussed in other reviews and papers (McAdams and Arkin, 1997; Vilar *et al.*, 2003). Here, to illustrate this process, let us consider a simple hypothetical genetic switch shown in Fig. 3. This system is bistable and resembles several generic genetic switches (Kepler and Elston, 2001; Ptashne, 1992). The reaction network considered here is exceedingly simple: proteins A and B can both bind to an operator site which in turn controls the expression of A and B. Both A and B dimerize, and when A_2 is bound to the operator, B is not transcribed. When B_2 is bound to the operator, A is not transcribed. Thus A and B mutually inhibit each other.

To analyze this system stochastically, we can generalize the discussion in the introduction and consider a set of reactions. The system is characterized by the number of A proteins (N_A), the number of B proteins (N_B), and the number of A and B dimers (N_{A2} and N_{B2}). The operator is either empty or docked by A_2 or B_2. A and B can also be degraded. Changes in these variables are described by stochastic jumps with transition rates given in Fig. 3. For example, N_A can increase by one from the transcription process with rate $5k$. However, this increase only occurs when there is an A_2 bound to the operator. N_A can decrease by one from degradation, with rate $0.25k$. N_A can also decrease by two from the formation of an A_2 dimer with rate $5k$. A similar analysis can be done for N_B. A_2 and B_2 can also bind to the operator with rate $5k$. It is then a matter of applying the stochastic simulation procedure in Section V to obtain stochastic trajectories such as in Fig. 3.

The result from such a stochastic simulation is that the system will randomly switch between the configuration where there are many more A than B and the configuration where there are many more B than A. Most of the time, the system is dominated by either A or B (Fig. 3). Switching probability is determined by a number of factors, including the rate of A and B production and the difference

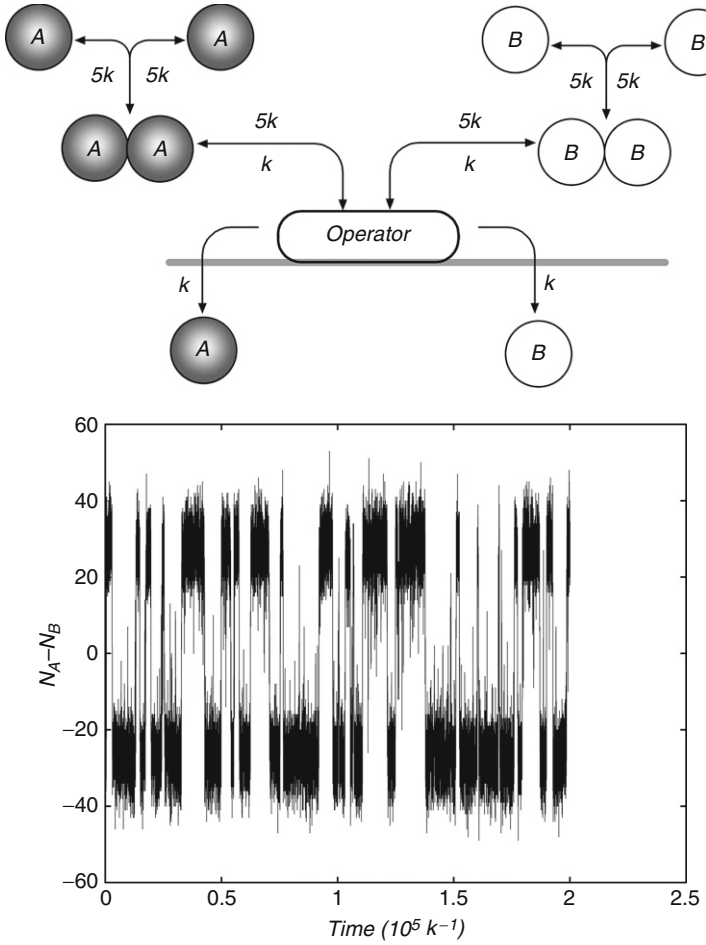

Fig. 3 A simple genetic switch. Here, A and B can form dimers. The dimers can bind to the operator to facilitate production of A or B. If the operator binds the dimer, only A is produced. If the operator binds the B dimer, only B is produced. The rate constants for these reactions are measured in units of k. The lower panel shows a typical stochastic simulation trajectory. The system switches mostly between A and B. The switch-like behavior would be absent if a kinetic model is used to describe the same system.

between N_A and N_B. Interestingly, if the same transcription cycle is analyzed using standard kinetic equations (namely a meanfield treatment), one will not observe the stochastic switching. This is because in the kinetic treatment, one computes the average probability of the operator being occupied by A_2 or B_2. This probability is roughly ½. This leads to simultaneous production of A and B with ½ of the rate, and the difference between N_A and N_B is zero at steady state. However, in the

stochastic model, the operator is either occupied by A_2 or B_2, leading to production of only one of the species. Therefore, the switch-like behavior of the operon only manifests itself when the stochastic model is considered. This is not a surprising result because we have noted that kinetic models are approximations to Markov models.

Beyond the simple genetic switch, any reaction system with any number of reacting partners can be analyzed using the stochastic approach. These are not limited to genetic switches, but include signaling, chemotactic, and developmental networks in bacteria and eukaryotic cells. However, there are a number of practical issues. Typically, the reaction rates *in vivo* are not known. In many cases, the reacting species and the relevant genes have not been identified. The discovery process of finding the important signaling pathways is usually the bottleneck. Cells also have redundant pathways, probably from evolutionary history in different environments. Stochastic models are useful in these problems to check the proposed mechanisms.

III. Molecular Motors and the Inclusion of Biomolecular Structure in Stochastic Models

Single molecule techniques can now measure conformational changes of biomolecules in real time. Conformational fluctuations are monitored using fluorescence resonance energy transfer (FRET), tethered optical traps, or direct fluorescent labeling of the protein (Bryant *et al.*, 2003; Myong *et al.*, 2005; Noji *et al.*, 1997; Visscher *et al.*, 1999; Yildiz *et al.*, 2003). For example, rotational motion of F_1-ATPase is first observed by attaching fluorescent F-actin to the rotating γ-subunit, and the observed motion is stochastic with occasional back rotation (Noji *et al.*, 1997). Stochastic stepping motions are observed in processive motors such as myosin-V (Mehta *et al.*, 1999), kinesin (Visscher *et al.*, 1999), and dynein (Mallik *et al.*, 2004) on F-actin and microtubules. All of these motors hydrolyze single ATP molecules as fuels. Interpretations of these experiments require stochastic models.

As we discussed above, stochastic models describe stochastic jumps between basins on an energy landscape. Therefore, the central question for molecular motors is: what is the energy landscape that explains the observed rotation or stepping behavior (Bustamante *et al.*, 2001)? How do we explain the observed step-size distributions and force–velocity curves? Since conformational fluctuations are measured in experiments, we need to consider a free energy surface as a function of the motor conformation. Since motors also bind and hydrolyze ATP, it is also necessary to consider the free energy surface as nucleotides are exchanged in the binding pocket. Thus, the general free energy surface is a function of two sets of variables, $V(q,\sigma)$, where q is the set of variables that describe the conformation of

the motor and σ is the set of variable that describe the nucleotide occupancy. In the case of F_1-ATPase, since the conformation is related to the rotation of the γ-subunit, q may be taken as θ, or the angle of the γ-subunit (Wang and Oster, 1998). (Note by taking this definition, we have assumed a tight-coupling between γ rotation and conformational changes in the α and β-subunits.) The nucleotide occupancy is determined by the occupancy states of the three catalytically active β-subunits; $\sigma = (s_1, s_2, s_3)$. Since each β-subunit can be either empty, ATP bound, ADP-Pi bound, and ADP bound, σ_i ranges from 1 to 4 and σ ranges from 1 to 64.

Free energy surfaces of biomolecules, in principle, can be computed from the known molecular structure. In practice, the computation for such large molecules is prohibitively costly. A coarse-grained approach is to consider the known conformational changes in the molecule and model them as simple mechanical elements. In F_1-ATPase, a hinge bending motion is seen in the β-subunit during the ATP hydrolysis cycle. The hinge motion is coupled to γ-subunit rotation through a linkage (Fig. 4A and 4B). Therefore, we can model the opening and closing of the β-subunit as a spring and map this motion to γ-subunit rotation (Fig. 4). The readers can refer to published articles for the details of $V(\theta,\sigma)$ (Sun *et al.*, 2004; Wang and Oster, 1998). This type of approach for specifying the energy surface does not include the atomic structure, but does incorporate the approximate large-scale structure of the protein. When possible, the mechanical constants may be computed from molecular dynamics simulations. Therefore, this approach potentially makes contact with several different kinds of data.

Note that in this formulation, the surface is continuous in the θ direction, but discontinuous or discrete in the σ direction. This is a simplification similar to going from Eq. (1) to (2). The energy surface, in reality, is continuous in all directions and there are barriers separating regions labeled by σ. The barriers are related to nucleotide binding and hydrolysis, and these atomic processes should depend on larger scale conformation of the protein. Instead of specifying the details of these energetic barriers in the σ direction, we describe the stochastic transitions between different σs by transition rates $k_{\sigma\to\sigma'}(\theta)$. The transition rate is determined by the barrier region of the energy landscape, which implies that the barriers are functions of protein conformation. Note that a connection to Bell's model of force-dependent reaction kinetics can be made here (Bell, 1978). Bell's model states that the reaction rate is modified exponentially in the presence of a mechanical force. This statement can be derived from a more detailed analysis using Kramer's rate theory (Hummer and Szabo, 2003). Here, in our context, if a force is applied to the motor protein, the protein will change its conformation, and the reaction rate changes. Reaction rates measured with purified proteins correspond to $k_{\sigma\to\sigma'}(\theta)$ for $\theta = \theta_0$, where θ_0 is the mechanical equilibrium conformation. Therefore, purified protein measurements can be used to estimate $k_{\sigma\to\sigma'}(\theta)$.

For processive molecular motors such as myosin-V, kinesin, and dynein, a similar analysis is also necessary to explain experimental measurements. Note that the situation is complicated by the fact that motors can bind and unbind from cytoskeletal tracks. The nucleotide occupancy variable, σ, needs to be

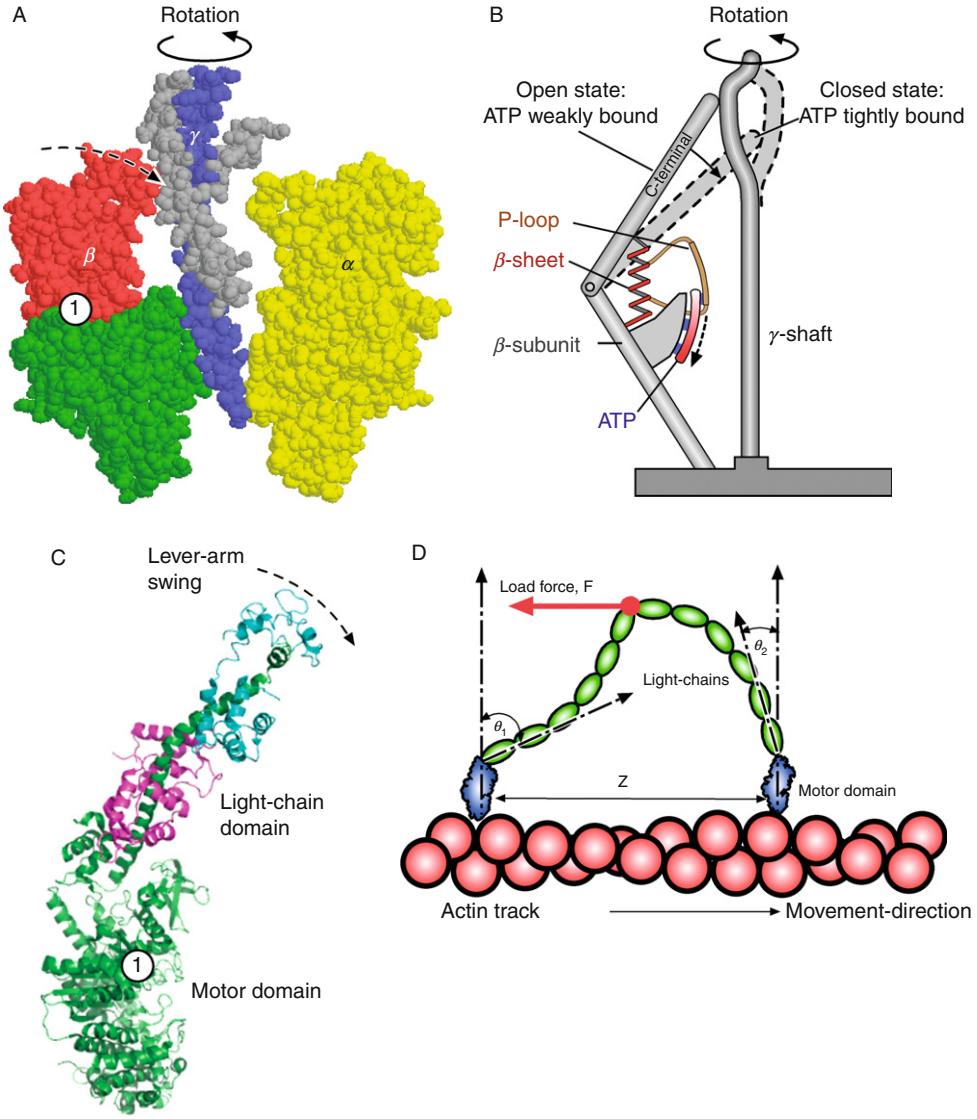

Fig. 4 Stochastic models of molecular motors incorporating the rough motor structure. (A) The γ-subunit of F_1-ATPase undergoes a rotation when the β-subunit is hydrolyzing ATP. A hinge bending motion in the β-subunit pushes the γ-subunit. The ATP binding site is labeled by 1. (B) To obtain an energy landscape description, we can consider the spring-like motion of the β-subunit and map it to the rotation of γ. (C) Myosin motors change conformation by changing the angle between the motor domain and the light-chain domain. (D) When two myosins are connected by the light-chain domain, the combined energy landscape can be written as Eq. (5). This energy depends on the conformations of the motor domains $(\theta_1, \phi_1, \theta_2, \phi_2)$, and the relative separations of the motor domains on F-actin, z. Any applied load force, F, also changes the motor conformation. Because actin binding sites are helical, binding to different sites changes the overall conformational energy. This leads to preferential binding to a few sites.

generalized to describe cytoskeleton binding of the motor domain. For the myosin-V motor, each motor domain can bind and hydrolyze ATP while attached or detached from F-actin. There are also two motor domains. Thus, $\sigma = (s_1, s_2)$ where $\sigma_{1,2}$ describes the nucleotide occupancy and F-actin binding activities of each motor. Also, unlike F_1-ATPase, conformational changes in the motor domains are not tightly coupled to the observed motor movement. Rather, the conformational changes of the motors change the conformation of the light-chain domain (also a coiled-coil motif) that links the motor domains. Examining the X-ray structures of the myosin motor domain shows that the conformational change is a rotation of the converter domain that reorients the light-chain domain (Fig. 4). The overall free energy surface can be written as (Lan and Sun, 2005, 2006).

$$V(\theta_1, \phi_1, \theta_2, \phi_2, z, \sigma_1, \sigma_2) = V_0(\theta_1, \phi_1, \sigma_1) + V_0(\theta_2, \phi_2, \sigma_2) + V_l(\theta_1, \phi_1, \theta_2, \phi_2, z) \quad (5)$$

where (θ_i, ϕ_i) is the 3-D orientation of the light-chain domain with respect to the motor (Fig. 4), z is the distance between the bound motor domains, and V_0 is the conformational energy of the monomer. And just as in the F_1-ATPase, V_0 depends on the occupancy of the nucleotide and F-actin binding activity. V_l is the conformation of the light-chain domain, and in addition to the conformational variables, it may also depend on any externally applied forces such as from an optical trap. Note that V_l is independent of (σ_1, σ_2), i.e., the light-chain domain does not have any catalytic activity. Again, simple mechanical elements can be used to model V_0 and V_l. We modeled V_0 as a mechanical spring and V_l as an elastic rod (Lan and Sun, 2005, 2006). This allows us to compute the energy surface with only a few parameters. It also allows us to compute the rough conformation of the molecule as it processes on F-actin.

The mechanical formulation outlined here predicts an asymmetrical structure when both motor domains are bound to F-actin and contain ADP. In this situation, the motor domains would like to have the same conformation. However, the light-chain domain connecting the motor domains resists such a conformation. If we compute the new mechanical equilibrium structure, the leading motor domain will have a different conformation than the trailing domain, and $(\theta_1, \phi_1) \neq (\theta_2, \phi_2)$. This suggests that ADP release rates from the bound motors will be different and the leading motor probably releases ADP slowly, resulting in preferential detachment of the trailing motor. This is a necessary outcome of the model to predict stepping behavior. Also, the free energy surface can be used to compute the energy difference between the situation where only one motor is bound and the other is free, and the situation where both motor domains are bound. This energy difference is a function of the relative distance between the bound motors, z. This energy difference also determines the relative binding rate of the free motor. The outcome determines the preferred binding site of the free motor. Since F-actin has a helical structure in its binding sites, the most favorable binding site occurs at $z = 36$ nm where the light-chain domain has the smallest deformation. These results are all consistent with experimental findings. Since the

approximate structures of the domains are included, it is possible to examine structural mutants. For example, truncating the light-chain domain leads to smaller steps, although the stepping behavior still persists (Lan and Sun, unpublished results). If the light-chain is made softer, such as in the myosin-VI isoform, the model can also predict the stepping behavior by changing the mechanical constants (Lan and Sun, 2006).

The same formalism outlined here is equally applicable for understanding other processive motors such as kinesin and dynein. The linkages between the motor domains in kinesin and dynein are structurally more complex and simple mechanical models are perhaps inadequate. Nevertheless, the overall energy of these motors can be decomposed in the same fashion as Eq. (5). If the structural changes associated with ATP hydrolysis and microtubule binding are known, kinesin and dynein can be understood using the framework outlined here.

We have shown that approximate biomolecular structure can be used to construct the free energy surface of molecular motors. Given the free energy surface, conformational motions of the motor can be computed using the stochastic equation of Eq. (1), or simplified schemes where only energetic basins and jumps between basins are considered in a Markovian stochastic model. The outcomes of these computations are essentially the same if the proper energy surface is considered first. Moreover, stochastic trajectories can be computed from these models and the results resemble many aspects of experimental data, and show features such as backward steps and substeps. Detailed structural explanations can be given for these features.

IV. Cytoskeleton and Cytoskeletal Network Structures

Stochastic models can be extended to consider the dynamic nature of the cytoskeletal network inside moving cells. The simplest situation is the growth of a single cytoskeleton filament such as F-actin. Here, the biochemistry of F-actin polymerization has been measured and quantified (Blanchoin, *et al.*, 2000; Marchand *et al.*, 1995; Mullins *et al.*, 1998; Pollard, 1986; Pollard and Borisy, 2003; Pollard *et al.*, 2000). Under typical cytoplasmic conditions, the barbed (plus) end of F-actin grows much faster than the pointed (minus) end. Thus, to simplify matters, let us neglect the pointed end and consider the growth dynamics at the barbed end. Let us also assume monomers are added in an alternating fashion, i.e., monomers are added to the left protofilament, then to the right, and so forth. These simplifications lead to the following equation for the probability of observing a filament with n-monomers:

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = k_+ P_{n-1} + k_- P_{n+1} - (k_+ + k_-) P_n \tag{6}$$

where $k_{\pm}$ are the monomer addition and subtraction rates. A filament of length n can occur by either adding a monomer from $n-1$, or losing a monomer from $n+1$. The monomer addition rate, k_+, will depend on the cytoplasmic monomer concentration. Equation (6) is similar to Eq. (3), and is also equivalent to a diffusion-convection equation for the probability of observing a filament with length l:

$$\frac{\partial P(l)}{\partial t} = D\frac{\partial^2 P}{\partial l^2} - v\frac{\partial P}{\partial l} \tag{7}$$

where $D = k_+\Delta^2$ is an effective diffusion constant and Δ the length increase with addition of one monomer, $v = (k_+ - k_-)\Delta$ is a "growth velocity." The diffusion–convection nature of Eq. (7) suggests that the length distribution will shift as well as spread in time. Equation (7) can be solved analytically, but a stochastic simulation algorithm can be implemented for Eq. (6), and the results will be equivalent. The stochastic simulation algorithm involves examining the current configuration (length n), and the two possible destination states (lengths $n + 1$ and $n - 1$), and carrying out the procedure in Section V.

Generalization to more than one filament is straightforward if the filaments are dynamically independent. The length distribution of a set of filaments is given by Eq. (7). Stochastic simulations of a set of filaments can be accomplished by recognizing that for any given configuration, there are many destination states because each filament can add or subtract monomers. If the filaments behave identically, then the rates of adding or subtracting monomers are identical. The only difference is that the number of accessible states is larger, i.e., $2N$ where N is the number of filaments.

In the cytoplasm, however, the growth of F-actin filaments is influenced by many factors (Pollard and Borisy, 2003), and different cell types exhibit a variety of behaviors. We will focus on the leading edge of motile cells such as the fish keratocyte which contains a dendritic network of actin filaments (Pollard and Borisy, 2003). A number of factors influence the network structure *in vivo*. First, actin monomer concentrations are not expected to be uniform throughout the cell; this influences the addition rate, k_+. Second, capping proteins and other actin-end proteins competes with monomer addition. Third, filaments physically interact with each other and can bundle and crosslink. Fourth, new filaments are initiated from mother filaments via the activated Arp2/3 complex, leading to a branched network structure. Final, filament severing proteins such as cofilin and profilin will disrupt networks and causes depolymerization. All of these factors contribute to the dynamic nature of the cellular leading edge. Biochemical evidences of these influences are summarized in excellent reviews (Pollard and Borisy, 2003). Additional questions regarding how extracellular signals are transmitted to growing filaments and the influences of the plasma membrane and membrane tethering proteins such as formin (Dickinson *et al.*, 2004; Kovar and Pollard, 2004) are still being debated.

Stochastic models of growing cytoskeletal networks can, in principle, incorporate all of the factors mentioned in the previous paragraph. However, the number of relevant Markov states would be enormous, and in many cases, the rate constants for modeling the stochastic jumps are still not known. Also, the mechanics and elastic properties of the cytoskeletal network also contribute to the dynamics of the leading edge. For these reasons, a full simulation of the cellular leading edge is presently unrealistic. However, simpler questions can be addressed using stochastic models. Simulations of growing networks have been used to examine the actin comet tail behind moving *Listeria monocytogenes*, and the actin networks behind rocketing beads (Alberts and Odell, 2004; Carlsson, 2001, 2003). Here, we will discuss the morphology of networks in the moving lamellipodium of fish keratocytes.

Electron microscopy studies of the lamellipodium in keratocytes show a dendritic branched structure for the F-actin network where the barbed ends of filaments are mostly pointed $\pm 35^\circ$ with respect to the leading edge (Maly and Borisy, 2001). From structural studies, it is known that the new daughter filament nucleated by the Arp2/3 complex forms 70° angle with the mother filament (Volkmann *et al.*, 2001). However, monomers in F-actin are arranged in a helical fashion and Arp2/3 can nucleate daughter filaments in all directions. When considering the 3-D nature of the network, it is unclear why the filaments are pointed in $\pm 35^\circ$ directions. Capping of F-actin ends are important in this situation because filaments away from the leading edge are probably quickly capped and do not grow further. The geometry of the network results from the competition between capping and branching. A stochastic simulation of this system, including branching, capping, and filament growth can examine the orientational distribution of filaments under different conditions. In this simulation, the geometry of new filaments is incorporated. The geometry of the plasma membrane is also included to examine the 3-D structure of the actin network (Atilgan *et al.*, 2005). The simulation shows that the filaments are pointed in $\pm 35^\circ$ directions only if the mother filament is pointed in a restricted set of directions (Fig. 5). Mathematically, this implies that the Arp2/3 branching rate is a function of the orientation of the mother filament (Atilgan *et al.*, 2005). This curious discovery can be explained if we consider an activated Arp2/3 complex that is attached to the plasma membrane, and the Arp2/3 is not freely diffusing in the cytoplasm (Fig. 5). New daughter filaments are generated only when the mother filament interacts with the attached Arp2/3 complex.

Several proteins are implicated in the activation of Arp2/3, and exactly how biochemical signals are translated to Arp2/3 is still being investigated. Additional questions can be posed about the localization of these signaling complexes and the effect of membrane curvature on signaling protein distribution (Atilgan and Sun, 2007; Atilgan *et al.*, 2005). Stochastic simulations studies, however, reveal that the geometry and the physical mechanism of filament growth are important in generating the observed network morphology. Further studies of F-actin dynamics *in vivo* can answer these questions.

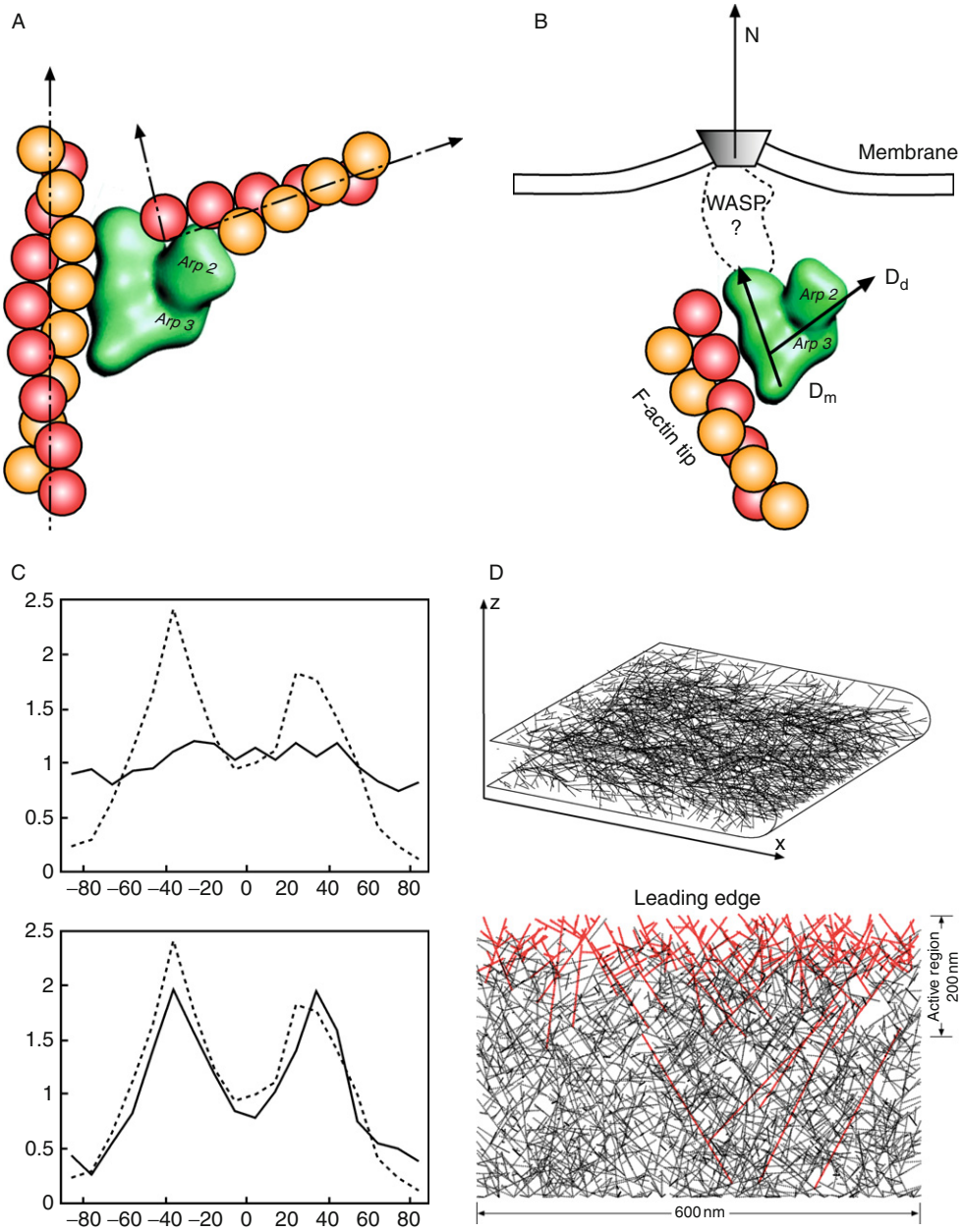

Fig. 5 Stochastic simulations can be used to analyze cytoskeletal network morphology in the lamellipodium. (A) New actin filaments are nucleated by Arp2/3. The structure of the filament junction is known, and is incorporated into the stochastic simulation. (B) If the Arp2/3 complex is still attached to the membrane during the nucleation of new filaments, then only filaments with the right orientation can generate new filaments. Here, $\mathbf{D}_m$ and $\mathbf{D}_d$ are the directional vectors of the mother and daughter

V. Procedures

In this section, the computational procedure for implementing Gillespie stochastic simulations is given. Before implementing the computer code, it is necessary to identify the variables that specify a particular Markov state. For example, when simulating a diffusing particle, a state would be specified by the spatial position of the particle on a grid. From any state, it is also necessary to identify all the possible states that are connected to it in one Markov jump. For a diffusing particle in 3-D, there are six neighboring states connected to any state. Finally, the transition rates to all the neighboring states, k_i, must be tabulated. The net escape rate is the sum of the transition rates to neighboring states: $K = \sum_{i=1}^{\text{neighbors}} k_i$. The simulation procedure can be listed as follows:

1. Establish an initial state of the system. This depends on the problem at hand. If we are modeling reactions, then an initial collection of reactants and products should constitute the initial state. If the system is in equilibrium, then the initial state can be randomly picked from the equilibrium distribution.
2. Compute the net escape rate from the current state, $K = \sum_{i=1}^{\text{neighbors}} k_i$.
3. Select a random number, $\triangle t$, according to the escape time distribution $e^{-K\Delta t}$. This is done by selecting a random number u distributed uniformly between 0 and 1, the escape time from the current state is $\Delta t = -(1/K) \ln u$.
4. Divide the interval from 0 to K into subintervals of length k_i, i.e., one subinterval for each transition rate to destination state i. Select another random number between 0 to K, v. Determine the subinterval j where v resides. This is the destination state of the transition event.
5. Update the current state to the destination state j, and record the jump time $\triangle t$. Obtain the current time by adding up the jump times.
6. Repeat step (2) for the new state.

This procedure generates a sequence of states and jump times, which we call a stochastic trajectory. From this sequence of states and times, it is possible to evaluate averages such as Eq. (4) by simply averaging over the trajectories. The most difficult step in this procedure is identifying all the possible neighboring states of any given state. For example, in the cytoskeleton simulation, all of the monomers in F-actin can potentially bind Arp2/3. Only a subset of these monomers can actually succeed in

filaments formed by the Arp2/3 junction, respectively. (C) The computed filament orientational distribution (solid lines) compared with the distribution from electron microscopy (dotted line). The upper panel shows the distribution when all filaments are allowed to generate new filaments. This corresponds to freely diffusing Arp2/3 in the cytoplasm. The lower panel shows the distribution when Arp2/3 is attached to the leading edge membrane. The lower panel results agree with experimental data. (D) Some representative F-actin network generated from stochastic simulations. Capping and branching events are included. The lower panel shows the projected network to the x–y plane. The red filaments are still growing, the black filaments have been capped.

binding. This is determined by the orientation of the mother filament. Therefore, during the simulation, the orientation of the mother filament is stored and the information is used to determine if an Arp2/3 can be added to the filament.

VI. Discussion and Concluding Remarks

In this review, we have examined the basic physics behind stochastic methods and applied the Gillespie method to several biological problems. The procedure to generate Gillespie stochastic trajectories is exceedingly simple and has wide applicability. We have discussed stochastic models for gene expression, stochastic movement of molecular motors, and the growth of cytoskeletal networks. We have shown that structural information can be incorporated into stochastic models to answer physical and morphological questions.

There are some technical issues when implementing stochastic simulations. We will comment on three of these issues. First of the issues has to do with rare events. The stochastic algorithm will generate the most probable stochastic trajectories, and the final state of the trajectory is usually a highly probable destination state. Suppose one is interested in a final state that is rarely visited by the trajectory. The algorithm does not offer a way to select for these rare trajectories. In order to obtain them, one has to simulate many trajectories, hoping for a few of them that reach the desired final state. Most of the computational effort is wasted in simulating undesirable but frequently occurring trajectories. To overcome this, one must develop an algorithm that generates trajectories with fixed end states. This is possible if one can compute the absolute probability of the stochastic trajectories explicitly. This problem has been analyzed and a possible solution is offered (Harland and Sun, 2007; Sun, 2006).

The second issue has to do with disparate time scales where some stochastic jumps occur at a much faster rate than others. During the simulation, most of the selected jumps will be fast jumps. In reaction–diffusion systems, this can be prevalent when the rate of diffusion is much faster than the rate of reaction. Most of the stochastic jumps will be diffusive steps. Of course, the simulation is reproducing the correct physics. But most of the computational time is wasted on generating trivial jumps. A possible solution to this is the τ-leap method (Gillespie, 2001). Instead of simulating these fast jumps explicitly, one can leap over a sequence of jumps to advance quickly.

The last issue has to do with system size. Stochastic simulations are computationally demanding when the number of Markov states is prohibitively large. In practice, the net escape rate from any state, K, is a large number so that the jump time, $\triangle t$, can be small. Many jumps are needed to simulate a long trajectory. Currently, there is a size limitation on how many reactions one can effectively simulate. Again, alternative algorithms may be devised for specialized problems or a completely new simulation algorithm may be devised to overcome the scaling problem.

Nevertheless, the ease of implementation makes stochastic models very attractive for examining complex systems ranging from single proteins to the cell. Simulations can be devised quickly to check the proposed mechanisms. Sometimes, simulations can raise additional questions and motivate new experiments. Thus, stochastic simulation methods should be a prominent tool for quantitative biologists working in both experimental and theoretical frontiers.

VII. Appendix Glossary of Terms

Several terms used in this chapter may be unfamiliar to the reader. A list of plain language definitions is given below.

Stochastic Models: A model that determines the configurations of a dynamical system probabilistically. This is in contrast with "dynamical models," which imply that given a configuration of system and laws governing the dynamics, future configurations are deterministic. Newtonian equation of motion is deterministic while the Langevin equation of Eq. (1) is stochastic.

Markov Models: A model based on a set of predefined states and a sequence of stochastic jumps between states. The stochastic jumps satisfy the Markov property which implies that the jumping probability only depends on the present state of the system, and does not depend on the history of states. Markov models are a subset of stochastic models.

Kinetic Models: A model to analyze changes in the average concentration of a set of reacting species. It usually assumes that the reactants are well-mixed and obey the law of mass action. Kinetic models are deterministic in the sense that given a set of initial concentrations and a reaction mechanism, all future concentrations are determined.

Importance Sampling (Metropolis Monte Carlo): A method to sample the most probable regions of a multi-dimensional probability distribution. A common application is in molecular dynamics where the Boltzmann distribution is a multi-dimensional probability distribution of atomic configurations. This method is usually applied for equilibrium simulations.

Gillespie Stochastic Simulation Method: A method to compute a sequence of stochastic states and jump times between states (or a stochastic trajectory). The method was discussed by Gillespie in the context of chemical reactions, although an equivalent method was mentioned by Bortz, Kalos, and Lebowitz (BLK) in the context of magnetic spin systems. Stochastic simulations of Markov models are usually based on this method. It can be shown that the Gillespie method is a form of importance sampling in trajectory space (see Sun *et al.*, 2004).

Mechanical Models: A model used to understand the dynamics of a system through the specification of forces between interacting components. The simplest mechanical model is an elastic spring which specifies the force, f, between two particles as proportional to the distance, x, between them: $f = -kx$. k is a spring constant or force constant.

Elastic Models: A subset of mechanical models which assumes that forces can be derived from a mechanical energy. The mechanical energy can be computed from atomistic models, or continuum theories with a few mechanical constants.

References

Alberts, J. B., and Odell, G. M. (2004). *In silico* reconstitution of *Listeria* propulsion exhibits nano-saltation. *PLoS Biol.* **2,** 2054–2066.

Atilgan, E., and Sun, S. X. (2007). Shape transitions in lipid membranes and protein mediated vesicle fusion and fission. *J. Chem. Phys.* **126,** 095102:1–10.

Atilgan, E., Wirtz, D., and Sun, S. X. (2005). Morphology of the lamellipodium and organization of actin filaments at the leading edge of crawling cells. *Biophys. J.* **89,** 3589–3602.

Bell, G. I. (1978). Models for the specific adhesion of cells to cells. *Science* **200,** 618–627.

Blanchoin, L., Pollard, T. D., and Mullins, R. D. (2000). Interactions of ADF/cofillin, Arp2/3 complex, capping protein and profilin in remodeling of branched actin filament networks. *Curr. Biol.* **10,** 1273–1282.

Bortz, A., Kalos, M., and Lebowitz, J. (1975). A new algorithm for Monte Carlo simulation of ising spin systems. *J. Comp. Phys.* **17,** 10–18.

Bryant, Z., Stone, M. D., Gore, J., Smith, S. B., Cozzarelli, N. R., and Bustamante, C. (2003). Structural transitions and elasticity from torque measurements on DNA. *Nature* **424,** 338–341.

Bustamante, C., Keller, D., and Oster, G. (2001). The physics of molecular motors. *Acc. Chem. Res.* **34,** 412–420.

Carlsson, A. E. (2001). Growth of branched actin networks against obstacles. *Biophys. J.* **81,** 1907–1923.

Carlsson, A. E. (2003). Growth velocities of branched actin networks. *Biophys. J.* **84,** 2907–2918.

Chandler, D. (1987). "Introduction to Modern Statistical Mechanics." Oxford University Press, Oxford.

Dickinson, R. B., Luzelena, C., and Purich, D. L. (2004). Force generation by cytoskeletal filament end-tracking proteins. *Biophys. J.* **87,** 2838–2854.

Frenkel, D., and Smit, B. (1996). "Understanding Molecular Simulations." Academic Press, New York.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81,** 2340–2361.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115,** 1716–1733.

Hanggi, P., Talkner, P., and Borkovec, M. (1990). Reaction-rate theory: Fifty years after Kramers. *Rev. Mod. Phys.* **62,** 251–341.

Harland, B., and Sun, S. X. (2007). Stochastic path sampling and the mean first passage time for Markovian dynamics. *J. Chem. Phys.* **127,** 104103.

Hummer, G., and Szabo, A. (2003). Kinetics from non-equilibrium single-molecule pulling experiments. *Biophys. J.* **85,** 5–15.

Kepler, T. B., and Elston, T. C. (2001). Stochasticity in transcriptional regulation: Origins, consequences and mathematical representations. *Biophys. J.* **81,** 3116–3136.

Kerr, R. A., Levine, H., Sejnowski, T. J., and Rappel, W. J. (2006). Division accuracy in a stochastic model of Min oscillations in *Escherichia coli*. *PNAS* **103,** 347–352.

Kovar, D. R., and Pollard, T. D. (2004). Insertional assembly of actin filament barbed ends in association with formins produces piconewton forces. *Proc. Natl. Acad. Sci. USA.* **101,** 14725–14730.

Lan, G., and Sun, S. X. (2005). Dynamics of myosin-V processivity. *Biophys. J.* **88,** 999–1008.

Lan, G., and Sun, S. X. (2006). Flexible light-chain and helical structure of F-actin explain the movement and step size of myosin-VI. *Biophys. J.* **91,** 4002–4013.

Lutkenhaus, J. (2007). Assembly dynamics of the bacterial minCDE system and spatial regulation of the Z ring. *Annu. Rev. Biochem.* **76,** 14.1–14.24.

Mallik, R., Carter, B. C., Lex, S. A., King, S. J., and Gross, S. P. (2004). Cytoplasmic dynein functions as a gear in response to load. *Nature* **427,** 649–652.

Maly, V. I., and Borisy, G. G. (2001). Self-organization of a propulsive actin network as an evolutionary process. *Proc. Natl. Acad. Sci. USA* **98,** 11324–11329.

Marchand, J.-B., Moreau, P., Paoletti, A., Cossart, P., Carlier, M.-F., and Pantoloni, D. (1995). Actin-based movement of Listeria monocytogenes: Actin assembly results from the local maintenance of the uncapped filament barbed ends at the bacterium surface. *J. Cell Biol.* **130,** 331–343.

McAdams, H. H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. *PNAS* **94,** 814–819.

Mehta, A., Rock, R., Rief, M., Spudich, J., Mooseker, M., and Cheney, R. (1999). Myosin-V is a processive actin-based motor. *Nature* **400,** 590–593.

Mullins, R. D., Heuser, J. A., and Pollard, T. D. (1998). The interaction of Arp2/3 complex with actin: Nucleation, high-affinity pointed end and capping, and formation of branching networks of filaments. *Proc. Natl. Acad. Sci. USA* **95,** 8161–8186.

Murray, J. D. (1989). "Mathematical Biology." Springer-Verlag, Heidelberg.

Myong, S., Rasnik, I., Joo, C., Lohman, T. M., and Ha, T. (2005). Repetitive shuttling of a motor protein on DNA. *Nature* **437,** 1321–1325.

Noji, H., Yasuda, R. M., Yoshida, M., and Kinosita, K. (1997). Direct observation of the rotation of F1-atpase. *Nature* **386,** 299–302.

Pollard, T. D. (1986). Rate constants for the reaction of ATP- and ADP-actin with the ends of actin filaments. *J. Cell Biol.* **103,** 2747–2754.

Pollard, T. D., Blanchoin, L., and Mullins, R. D. (2000). Molecular mechanisms controlling actin-based filament dynamics in muscle cells. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 545–576.

Pollard, T. D., and Borisy, G. G. (2003). Cellular motility driven by actin assembly and disassembly of actin filaments. *Cell* **112,** 453–465.

Ptashne, M. (1992). "A Genetic Switch: Gene Control and Phage λ." Blackwell, Oxford.

Reichl, L. E. (1998). "A Modern Course in Statistical Physics." Wiley, Chichester.

Risken, H. (1989). "The Fokker-Planck Equation." Springer-Verlag, Berlin.

Sun, S. X. (2006). Path summation formulation of the Master equation. *Phys. Rev. Lett.* **96,** 210602:1–4.

Sun, S. X., Wang, H., and Oster, G. (2004). Asymmetry in F1-atpase and its implications for the rotational cycle. *Biophys. J.* **86,** 1373–1384.

Vilar, J. M. G., Guet, C. C., and Leibler, S. (2003). Modeling network dynamics: The lac operon, a case study. *J. Cell Biol.* **161,** 471–476.

Visscher, K., Schnitzer, M. J., and Block, S. M. (1999). Single kinesin molecules studied with a molecular force clamp. *Nature* **400,** 184–189.

Volkmann, N., Amann, K. J., Stoilova-McPhie, S., Egile, C., Winter, D. C., Hazelwood, L., Heuser, J. E., Li, R., Pollard, T. D., and Hanein, D. (2001). Structure of Arp2/3 complex in its activated state and in actin filament branch junctions. *Science* **28,** 2456–2458.

Wang, H., and Oster, G. (1998). Energy transduction in F1 motor of ATP synthase. *Nature* **396,** 279–282.

Wang, H., Peskin, C. S., and Elston, T. C. (2003). A robust numerical algorithm for studying biomolecular transport processes. *J. Theor. Biol.* **221,** 491–511.

Yildiz, A., Forkey, J., McKinney, S. A., Ha, T., Goldman, Y. E., Selvin, P. R., and Myosin, V. (2003). Walks hand-over-hand: Single fluorophore imaging with 1.5 nm localization. *Science* **300,** 2061–2065.