Statistics and Related Topics in Single-Molecule Biophysics

Hong Qian[1] and S.C. Kou[2]

[1]Department of Applied Mathematics, University of Washington, Seattle, Washington 98195; email: hqian@u.washington.edu

[2]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138; email: kou@stat.harvard.edu

Annu. Rev. Stat. Appl. 2014. 1:465–92

First published online as a Review in Advance on November 4, 2013.

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-022513-115535

Keywords

diffusion, entropic force, fluctuation, Markov process, statistical inference, stochastic modeling

Abstract

Since the universal acceptance of atoms and molecules as the fundamental constituents of matter in the early-twentieth century, molecular physics, chemistry, and molecular biology have all experienced major theoretical breakthroughs. Although researchers had to wait until the 1970s to see individual biological macromolecules one at a time in action, the field of single-molecule biophysics has witnessed extensive growth in both experiments and theory since then. A distinct feature of single-molecule biophysics is that the motions and interactions of molecules as well as the transformation of molecular species are necessarily described in the language of stochastic processes, whether one investigates equilibrium or far-from-equilibrium living behavior. For laboratory measurements following a biological process, analysis of experimental data obtained by sampling individual participating molecules over time naturally calls for the inference of stochastic processes. The theoretical and experimental developments of single-molecule biophysics thus present interesting questions and unique opportunities for applied statisticians and probabilists. In this article, we review some important statistical developments in connection to single-molecule biophysics, emphasizing the application of stochastic-process theory and the statistical questions arising from modeling and analyzing experimental data.

1. INTRODUCTION

Although the concept of atoms and molecules can be traced back to ancient Greece, the corpuscular nature of atoms was firmly established only in the beginning of the twentieth century. The stochastic movement of molecules and colloidal particles in aqueous solutions, known as Brownian motion, explained by the diffusion theory of A. Einstein (1905) and M. von Smoluchowski (1906) and the stochastic differential equation of P. Langevin (1908)—confirmed experimentally through the statistical measurements of J.-B. Perrin (1912), T. Svedberg, and A.F. Westgren (1915)—played a decisive role in its acceptance (Perrin 1916). The literature on this subject is enormous. We refer readers to the excellent edited volume by N. Wax (1954), which includes now classical papers by Chandrasekhar, Uhlenbeck-Ornstein, Wang-Uhlenbeck, Rice, Kac, and Doob, and to M. Kac (1959) for a collection of lectures by one of the founding members of modern probability theory (Kac 1985).

Although physicists, ever since Isaac Newton, have been interested in the position and velocity of particle movements, chemists have always perceived molecular reactions as discrete events, even though no one had seen such until the 1970s. Two landmark papers that marked the beginning of statistical theories in chemistry (at least in the West) appeared in the 1940s (Kramers 1940, Delbrück 1940). First, Kramers (1940) elucidated the emergence of a discrete chemical transition in terms of a continuous "Brownian motion in a molecular force field" with two stable equilibria separated by an energy saddle and derived an asymptotic formula for the reaction rate. Probabilistically speaking, this formula corresponds to the rate of an elementary chemical reaction as a rare event (Schuss 2010). Second, Delbrück (1940) assumed discrete transitions with exponential waiting time for each and every chemical reaction and outlined a stochastic multidimensional birth-and-death process for a chemical reaction system with multiple reacting chemical species. Together, these two mathematical theories have established a path from physics to cell biology by (*a*) bridging the atomic physics with individual chemical reactions in aqueous solutions and (*b*) connecting coupled chemical reactions with dynamic chemical/biochemical systems. In 1977, Gillespie independently discovered Delbrück's chemical master equation approach (McQuarrie 1967) in terms of its Markovian trajectories on the basis of a computational sampling algorithm that now bears his name within the biochemistry community (Gillespie 1977). In fact, the simulation method can be traced back to Doob (1942).

Experimental techniques have experienced major breakthroughs in parallel with these theoretical developments. Perrin's investigations on Brownian motion gave perhaps the first set of single-particle measurements with stochastic trajectory. In the 1910s, spatial and temporal resolutions were on the order of micrometer and tens of a second. By the late 1980s, they became nanometer and tens of a millisecond. The observation of discrete stochastic transitions between different states of a single molecule was first achieved in the 1970s on ion channels, proteins imbedded in the biological cell membrane. This was made possible by the invention of the patch-clamp technique, together with exquisite electronics, for measuring small electrical currents (Sakmann & Neher 2009). To measure the stochastic dynamics of a tumbling single molecule in an aqueous solution, one must be able to see the molecule under a microscope for a sufficiently long time. For this purpose, an experimental technique to immobilize a molecule and highly sensitive optical microscopy are needed. These challenges were first overcome for enzyme molecules at room temperature in 1998 (Lu et al. 1998).

To statisticians and probabilists, it is clear that biophysical dynamics at the molecular level are stochastic processes. Thus, to characterize such dynamics, called fluctuations within the chemical physics literature, one needs stochastic models. In an experiment, if such processes are sampled over time, one molecule at a time, then the analysis of experimental data naturally calls for the

inference of stochastic processes. Therefore, the theoretical and experimental developments of single-molecule biophysics present great opportunities for applied statistics and probability.

The aim of this article is to review some important statistical developments in single-molecule biophysics from the construction of theoretical models to advances in experiments, mostly drawing from our own limited research experience. The discussion is far from complete, as the field of single-molecule biophysics, with a substantial background, is advancing too rapidly to be captured by a short review. Still, we hope to convey a certain amount of historical continuity as well as current excitement at the research interface between statistics and molecular biophysics. Special attention is paid to the application of stochastic-process theory and the statistical questions arising from the analysis of experimental data.

In this presentation, we discuss the underlying theory, experiments, and analysis of experimental data. The discussion of theory focuses more on the application of stochastic processes in modeling various problems in single-molecule biophysics, whereas the discussion of experiments and data focuses more on the statistical analysis of data. However, we want to emphasize that, similar to advances in modern science, theory and experiment/data go hand in hand: Development in one stimulates and inspires development in the other.

2. BROWNIAN MOTION AND DIFFUSION OF BIOLOGICAL MACROMOLECULES

Before we discuss Brownian motion and its profound implications in biophysics, we want to clarify some terminology because the term Brownian motion has different meanings when used in physics and chemistry versus when used in probability and statistics: For physicists and chemists, Brownian motion corresponds to the integral of the Ornstein–Uhlenbeck process (as discussed below); by contrast, for statisticians and probabilists, Brownian motion refers to the Wiener process, although both referents share the characteristic of $E[x^2(t)] \propto t$ for large t. Likewise, diffusion has a different meaning in statistics versus biophysics. In statistics and probability, diffusion processes typically refer to continuous-time and continuous-space Markov processes, such as Itō's diffusions. In biophysics, diffusion typically refers to the physical motion of a particle without an external potential; when there is drift, it is often called biased diffusion.

To facilitate our discussion, let us first review the derivation of the law of physical Brownian motion (Schuss 2010). Suppose we have a particle with mass m suspended in a fluid. Then according to Newton's equation of motion formulated by Langevin, the velocity $v(t)$ of the particle satisfies

$$m\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\zeta v(t) + F(t), \qquad 1.$$

where ζ is the damping coefficient and $F(t)$ is white noise—formally, the derivative of the Wiener process. To correctly represent an inert particle in thermal equilibrium with the fluid, the Langevin equation must have an important physical constraint that links the damping coefficient ζ with the noise level, because both the movement of the particle and the friction originate from one source—the collision between the particle and surrounding fluid molecules:

$$E[F(t)F(s)] = 2\zeta k_{\mathrm{B}} T \cdot \delta(t-s), \qquad 2.$$

where $\delta(\cdot)$ is Dirac's delta function, k_{B} is the Boltzmann constant, and T is the underlying temperature. Equation 2 is a consequence of the fluctuation-dissipation theorem for inert biophysical systems in statistical mechanics (Chandler 1987). Probabilistically speaking, a Markov-process model for an inert system that tends toward thermal equilibrium is necessarily reversible (Qian 2001, Qian et al. 2002).

In more rigorous probability notation, Equations 1 and 2 translate to

$$m\,\mathrm{d}v(t) = -\zeta v(t)\mathrm{d}t + \sqrt{2\zeta k_\mathrm{B}T}\;\mathrm{d}B(t), \qquad 3.$$

where $B(t)$ is the Wiener process and the formal association of $F(t) = \sqrt{2\zeta k_\mathrm{B}T}\;\mathrm{d}B(t)/\mathrm{d}t$ is recognized. The stationary solution of Equation 3 is the Ornstein–Uhlenbeck process (Wax 1954), which is Gaussian with mean function $E[v(t)] = 0$ and covariance function $E[v(t)v(s)] = (k_\mathrm{B}T/m)\exp\left(-\frac{\zeta}{m}|t-s|\right)$. Thus, for the displacement, $x(t) = \int_0^t v(s)\mathrm{d}s$, which can be recorded in single-particle tracking (SPT), its mean squared is

$$\begin{aligned} E[x^2(t)] = \mathrm{Var}[x(t)] &= \int_0^t \int_0^t E[v(s)v(u)]\mathrm{d}u\,\mathrm{d}s \\ &= 2\left(\frac{k_\mathrm{B}T}{\zeta}\right)t - 2\left(\frac{k_\mathrm{B}Tm}{\zeta^2}\right)\left(1 - e^{-\frac{\zeta}{m}t}\right). \end{aligned}$$

Therefore,

$$E[x^2(t)] \sim \left(\frac{k_\mathrm{B}T}{m}\right)t^2, \quad \text{for small } t; \qquad 4a.$$

$$E[x^2(t)] \sim 2\left(\frac{k_\mathrm{B}T}{\zeta}\right)t, \quad \text{for large } t. \qquad 4b.$$

Equation 4b gives the famous Einstein–Smoluchowski relation, which links the diffusion constant D with the damping ζ of the particle: $D = k_\mathrm{B}T/\zeta$. This equation is historically highly significant: By combining it with Stokes's law ($\zeta = 6\pi\eta r$) and the definition of the Boltzmann constant ($k_\mathrm{B} = R/N$), one obtains

$$D = \frac{RT}{6\pi\eta r N}, \qquad 5.$$

where η is the viscosity, r is the radius of the spherical particle, R is the gas constant, and N is the Avogadro constant.

An immediate experimental consequence of Equation 5 is that, by measuring the diffusion constant of a spherical particle, one can estimate the Avogadro constant! Indeed, experiments on Brownian motions have had a shining history in both physics and chemistry. For example, in 1926, Perrin and Svedberg won Nobel Prizes in physics and chemistry, respectively. Perrin studied the trajectories of Brownian motions, verifying Einstein's description of Brownian motion and providing one of the first modern estimates of the Avogadro constant, whereas Svedberg developed the method of analytical ultracentrifugation, using counts of Brownian particles in a well-defined volume and studying how this counting process evolves over time. Per Kac (1959), this counting process is referred to as the Smoluchowski process. Observations by both Perrin and Svedberg were performed on large colloids; nearly half a century would pass before such measurements were performed on biological macromolecules. As a version of the Svedberg experiment, fluorescence correlation spectroscopy (FCS) appeared in the 1970s (see Section 4), and the measurement of a single trajectory using the principle of spatial high-resolution by centroid localization (termed single-particle tracking, SPT) was developed in the 1980s. Notably, SPT is responsible for driving many recent advances in single-molecule biophysics and super-resolution imaging.

For experimental data from a true Brownian motion, a natural statistical aim is to obtain estimates of the diffusion constant. If the data consist of the trajectories of individual particles as in SPT, the diffusion constant can be estimated by either a least-square regression or a maximum likelihood estimate (MLE) (for a detailed discussion, see Section 2.1). If the data consist of particle counts over time, the statistical estimation becomes more involved (for a discussion starting with

the Smoluchowski process, which is non-Markovian, see Section 3) (Ruben 1963, McDunnough 1978).

In addition to estimating the diffusion constant, the experimental objective often is to investigate the motion that deviates from a simple Brownian motion. This aim has yielded many developments in statistical treatments of these data, prompting additional questions: What if there is a drift, if the space is not homogeneous, if the Brownian particles can reversibly attach to other stationary or moving objects, or if the particles are interacting (e.g., not independent)? With the emergence of super-resolution imaging, these questions are still being asked in laboratories; a systematic statistical treatment of these problems has yet to be developed (Weber et al. 2012).

2.1. Single-Particle Tracking (SPT) of Biological Molecules

Since the late 1980s, camera-based SPT has been a popular tool for studying the microscopic behavior of individual molecules (Saxton & Jacobson 1997). In such experiments, the trajectory of an individual particle is typically recorded through a microscope by a digital camera; the speed of the camera can be as fast as a few milliseconds per frame. The superb spatial resolution is due to centroid localization.

One of the most common statistical issues is to determine the diffusion constant D of the underlying particle from the experimental trajectory. If we denote $(x(t_1), \ldots, x(t_n))$ as the true positions of the particle at times $t_1, \ldots, t_n$, where $\Delta t \equiv t_i - t_{i-1}$ is the time interval between successive positions, then the experimental observations $(y_1, y_2, \ldots, y_n)$ are $y_i = x(t_i) + \varepsilon_i$, where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ are the localization (measurement) errors. If the particle's motion is Brownian, then, as shown in Equation 4b, the process $x(t)$ can be well approximated by $\sqrt{2D}B(t)$, where $B(t)$ is the standard Wiener process, provided $t \gg m/\zeta$. This leads to

$$y_i = \sqrt{2D}B(t_i) + \varepsilon_i. \qquad 6.$$

An intuitive estimate of D used by many experimentalists utilizes the mean square displacements (MSD) (H. Qian et al. 1991), such as

$$\hat{\rho}_k = \frac{1}{2(n-k)k\Delta t} \sum_{i=1}^{n-k} (y_{i+k} - y_i)^2, \quad k = 1, 2, \ldots,$$

which are averages of correlated (square) increments, or

$$\hat{\rho}'_k = \frac{1}{2\lfloor n/k \rfloor k\Delta t} \sum_{i=1}^{\lfloor n/k \rfloor} (y_{ik} - y_{1+(i-1)k})^2, \quad k = 1, 2, \ldots,$$

which are averages of nonoverlapping (square) increments. One can also try to combine them, for example, by weighting or a regression (against k) (Michalet 2010).

Given the parametric specification (Equation 6), another natural estimate of D is the MLE (Berglund 2010). Notably, the MLE and the optimal estimate based on MSD have comparable accuracy (Michalet & Berglund 2012). Furthermore, the estimation error in D decreases with n, the sample size (the number of camera frames), at the rather slow rate of $O(n^{-1/4})$, which contrasts with the familiar rate of $O(n^{-1/2})$ as in the central limit theorem (Gloter & Jacod 2001a,b; Cai et al. 2010).

The determination of the diffusion constant D serves many purposes, including estimating Avogadro's constant (Perrin's original aim), testing whether the underlying motion is Brownian, and elucidating detailed molecular mechanisms. For example, Blainey et al. (2009) studied how DNA-binding proteins move along DNA segments. If a DNA-binding protein simply slides along

the DNA, then the protein executes simple one-dimensional translational movements parallel to the DNA without rotation. By contrast, if a DNA-binding protein moves along the DNA through a helical path, then it retains a specific orientation with respect to the DNA helix and rotates with the helix (in a spiral fashion) (Halford & Marko 2004, Slutsky & Mirny 2004). If we measure a protein's position along the DNA over time, then the two motions are subject to different expressions of the diffusion constant: In the parallel motion, the diffusion constant is

$$D = \frac{k_B T}{6\pi\eta r},$$

as shown in Equation 5, where η is the viscosity and r is the size of the protein. In the helical motion, the diffusion constant is

$$D = \frac{k_B T}{\pi\eta r}\left[6 + \left(\frac{2\pi}{b}\right)^2 \left(8r^2 + 6r_{oc}^2\right)\right]^{-1}, \qquad 7.$$

where r_{oc} is the distance between the protein's center of mass and the axis of the DNA and b is the distance along the DNA traveled by the protein per helical turn. Equation 7 is derived from hydrodynamic considerations (Schurr 1979, Bagchi et al. 2008). The parallel and helical motions can thus be differentiated from the experimentally estimated diffusion constant. By tracking DNA-binding proteins with various sizes from different functional groups and estimating their diffusion constants from single-molecule experimental data, Blainey et al. (2009) found that the helical motion is the general mechanism.

2.2. Subdiffusion

As shown in Equation 4b, a key characteristic of Brownian motion is that the MSD $E[x^2(t)] \propto t$ for moderate and large t. In some physical and biological systems (Bouchaud & Georges 1990, Klafter et al. 1996), the motion is observed to follow $E[x^2(t)] \propto t^\alpha$ with $0 < \alpha < 1$. These motions are referred to as subdiffusion because $\alpha < 1$. One theoretical approach to model subdiffusion is to employ fractional calculus (such as the use of fractional derivatives) (reviewed in Metzler & Klafter 2000). We review an alternative approach here: a generalized Langevin equation (GLE) with fractional Gaussian noise as postulated in Kou & Xie (2004).

We start with a GLE (Chandler 1987),

$$m\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\zeta\int_{-\infty}^{t} v(u)K(t-u)\mathrm{d}u + G(t), \qquad 8.$$

where (*a*) a noise $G(t)$ having memory replaces the white noise and (*b*) the memory kernel K convoluted with the velocity make the process non-Markovian, contra the Langevin equation (Equation 1). Owing to the fluctuation-dissipation theorem, the memory kernel $K(t)$ and the noise are linked by

$$E[G(t)G(s)] = k_B T\zeta \cdot K(t-s)$$

(Zwanzig 2001). Note that the GLE reduces to the Langevin equation when K is the delta function.

Within the GLE framework, we are looking for a kernel function that can give subdiffusion. Because the white noise is the formal derivative of a Wiener process, which is the unique process that satisfies (*a*) being Gaussian, (*b*) having independent increment, (*c*) having stationary increment, and (*d*) being self-similar, a good candidate to generalize the Wiener process is a process with three properties: (*a*) Gaussian, (*b*) stationary, and (*c*) self-similar. It has been shown that the only class of processes that embodies all three properties is the fractional Brownian motion (fBm) $B_H(t)$ (Embrechts & Maejima 2002, Qian 2003), which has mean $E[B_H(t)] = 0$ and covariance

$E[B_H(t)B_H(s)] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right)$. $H \in [0,1]$ is called the Hurst parameter. $B_H(t)$ reduces to the Wiener process when $H = 1/2$.

Taking $G(t)$ in Equation 8 to be the (formal) derivative of fBm, $F_H(t) = \sqrt{2\zeta k_B T}\,\mathrm{d}B_H(t)/\mathrm{d}t$, we reach the model $m\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\zeta\int_{-\infty}^{t} v(u)K_H(t-u)\mathrm{d}u + F_H(t)$, where the kernel $K_H(t)$ is given by

$$K_H(t) = E[F_H(0)F_H(t)]/(k_B T\,\zeta) = 2H(2H-1)|t|^{2H-2}, \quad \text{for } t \neq 0. \qquad 9.$$

$F_H(t)$ is known as the fractional Gaussian noise (fGn).

In more rigorous probability notation, the model can be written as

$$m\,\mathrm{d}v(t) = -\zeta\left(\int_{-\infty}^{t} v(u)K_H(t-u)\mathrm{d}u\right)\mathrm{d}t + \sqrt{2\zeta k_B T}\;\mathrm{d}B_H(t). \qquad 10.$$

This equation is non-Markovian. Nevertheless, it can be solved in closed form via Fourier analysis (Kou 2008a). The solution $v(t)$ is a stationary Gaussian process, and the displacement $x(t) = \int_0^t v(s)\mathrm{d}s$ satisfies

$$E[x(t)^2] = \mathrm{Var}[x(t)] \sim \frac{k_B T}{\zeta}\left(\frac{2\sin(2H\pi)}{\pi H(2H-1)(2H-2)}\right)t^{2-2H} \propto t^{2-2H},$$

for large t. Therefore, the model with $H > 1/2$ leads to subdiffusion.

If there exists an external potential $U(x)$, $-U'(x(t))$ will be added to the right-hand side of Equation 8, yielding

$$\begin{aligned} \mathrm{d}x(t) &= v(t)\mathrm{d}t \\ m\,\mathrm{d}v(t) &= -\zeta\left(\int_{-\infty}^{t} v(u)K_H(t-u)\mathrm{d}u\right)\mathrm{d}t - U'(x(t))\mathrm{d}t + \sqrt{2\zeta k_B T}\,\mathrm{d}B_H(t). \end{aligned} \qquad 11.$$

For a harmonic potential $U(x) = \frac{1}{2}m\psi x^2$, the model can again be solved by the Fourier transform method (Kou 2008a).

Subdiffusive motion is observed in single-molecule experiments on protein conformational fluctuation (Yang et al. 2003, Min et al. 2005b) studying conformation fluctuation through the fluorescence lifetime of a protein. The fluorescence lifetime is a sensitive indicator, given its exponential dependence on the three-dimensional atomic arrangements of a protein. The stochastic fluctuation of the fluorescence lifetime, recorded in experiments, reveals the stochastic fluctuation in a protein's conformation. Detailed analysis of the autocorrelation function and three- and four-step high-order correlations of the experimental fluorescence lifetime data shows that (*a*) the conformation fluctuations of the two protein systems undergo subdiffusion, (*b*) the memory kernel is well described by Equation 9, (*c*) the conformation fluctuation is reversible in time, and (*d*) a harmonic potential captures the fluctuation quite well. These subdiffusive observations, therefore, directly support the notation of fluctuating enzymes, also known as dynamic disorder: As an enzyme molecule spontaneously changes its conformation, its catalytic rate does not hold constant. The different conformations of an enzyme molecule and their intertransitions thus could have direct implications on the enzyme's catalytic behavior (Min et al. 2005a) (see Section 5.5). From a pure statistics standpoint, inference and testing of the subdiffusive models beyond the autocorrelation function and the three- and four-step correlations are open issues.

3. PARTICLE COUNTING

The idea of counting the number of particles in a fixed region and using the temporal correlation of the resulting counting process to extract the kinetic parameters of the underlying experimental system has a long history, dating back to Smoluchowski's investigation of Brownian motion in

the early-twentieth century. Suppose we have indistinguishable particles, each undergoing independent Brownian motion. Let $n(t)$ be the number of particles at time t in a region Ω (such as an area illuminated under a laser beam). This counting process $\{n(t), t \geq 0\}$ is referred to as the Smoluchowski process. Under the assumption that the initial positions of the particles are uniformly distributed in a volume S (which is typically much larger than Ω), it can be shown that $E(n(t)) = |\Omega| / |S|$, and that for $t \gg m/\zeta$,

$$\mathrm{Cov}(n(t), n(t+\tau)) = \frac{|\Omega|}{|S|}\left\{1 - \frac{1}{(4\pi D\tau)^{3/2}}\int\int_{\mathbf{x}_1,\mathbf{x}_2\in\Omega} \exp\left(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||^2}{4D\tau}\right) d\mathbf{x}_1 d\mathbf{x}_2\right\}, \qquad 12.$$

where $|\Omega|$ and $|S|$ are the volumes of Ω and S, respectively, and D is the diffusion constant (Kac 1959, Ruben 1964, Brenner et al. 1978, McDunnough 1978, Bingham & Dunham 1997). Note that under $t \gg m/\zeta$, the Brownian diffusion is well approximated by the Wiener process, which is the basis for Equation 12. Historically, this result allowed the Brownian diffusion theory to be tested by particle counting, as Svedberg and Westgren did in the 1910s. It also allowed Smoluchowski to successfully account for the apparent paradox between the microscopic reversibility of an individual molecule's motion and the macroscopic irreversibility as in the second law of thermodynamics (Chandrasekhar 1943). Finally, it offers an experimental way to determine the diffusion constant.

Estimating D from experimental observations $(n(t_1), \ldots, n(t_M))$, where $\Delta t \equiv t_i - t_{i-1}$, is again a statistical issue. An intuitive method is to match the theoretical covariance function with the empirical one (Ruben 1964):

$$\frac{1}{M-1}\sum_{i=2}^{M}(n(t_i) - n(t_{i-1}))^2 = C(\Delta t, D), \qquad 13.$$

where $C(\Delta t, D)$ is the right-hand side of Equation 12, which is a function of Δt and D. The solution $\hat{D}$ of the generalized difference (Equation 13) is the estimate of D. Alternatively, one can also match lag-k square difference

$$\widehat{Cov}(k) := \frac{1}{M-k}\sum_{i=k+1}^{M}(n(t_i) - n(t_{i-k}))^2 = C(k\Delta t, D)$$

or use the nonlinear least square

$$\arg\min_D \sum_k \left(\widehat{Cov}(k) - C(k\Delta t, D)\right)^2$$

or its (weighted) variation to estimate D (Brenner et al. 1978).

Use of MLE to estimate D encounters the difficulties that the Smoluchowski process is non-Markovian and that it does not have analytically tractable joint probability function. Approximating the Smoluchowski process by an emigration-immigration (birth-death) process, which is Markovian, has been proposed (Ruben 1963, McDunnough 1978), where the birth and death rates can be set by making sure that the emigration-immigration and Smoluchowski processes share the same mean and covariance (for small Δt). Systematic comparison between the two different estimation methods—the one based on the empirical autocovariance function versus the quasi-likelihood estimate based on the emigration-immigration approximation—is an open issue.

The scheme of counting particles and utilizing the temporal correlation to extract kinetic parameters was further developed into FCS in the 1970s (see Section 4). With FCS, however, instead of exact counts, the fluorescence level of the underlying system, which depends on the molecules' counts, is recorded. The autocorrelation of the stochastic fluorescence reading can be used to estimate parameters such as the diffusion constant and the reaction rate.

4. FLUORESCENCE CORRELATION SPECTROSCOPY (FCS) AND CONCENTRATION FLUCTUATIONS

With the development of laser-based microscopy, the number of molecules in a very small region within an aqueous solution can now be measured and counted: Counting is based on the fluorescent light emitted from the molecules. If the molecules are continuously emitting fluorescence, then the measurement of stationary fluorescence fluctuation from a small region provides information on number fluctuation. Because fluorescent emission requires excitation of an incoming light, the small region is naturally defined by the laser intensity function $I(\mathbf{r})$, where $\mathbf{r} = (x, y, z)$ is the three-dimensional location of the particle (Rigler & Elson 2001); $I(\mathbf{r})$ can often be nicely represented by a Gaussian function $I(\mathbf{r}) = I_0 \exp(-2(x^2 + y^2)/\omega^2 - 2z^2/\omega_z^2)$.

For a collection of free-moving, identical, independent fluorescence-emitting particles, the theory is built on the function of a single Brownian motion: $I(X_t)$, where X_t is a three-dimensional Brownian motion, with diffusion coefficient D, confined in a large finite volume Ω. For comparison with a real experiment, we consider N i.i.d. Brownian motions and let $N, \Omega \to \infty$ such that $N/|\Omega| = c$ corresponds to the concentration of the particles in the real experiment (Qian et al. 1999), with $|\Omega|$ denoting the volume of Ω. Then the autocovariance function of $I(X_t)$ can be derived (Rigler & Elson 2001):

$$\mathrm{Cov}[I(X_{t+\tau}), I(X_t)] = \frac{var[I(X_t)]}{(1 + 4D\tau/\omega^2)(1 + 4D\tau/\omega_z^2)^{1/2}},$$

which can be used to obtain the diffusion constant D. This exact result and the corresponding experiments were developed in the 1970s. If the number of fluorescent particles is very large, then the measured stationary intensity $I(t)$ is essentially a Gaussian process with the mean and variance given by

$$\begin{aligned} E[I(t)] &= c \int_{\mathbb{R}^3} I(\mathbf{r})\mathrm{d}\mathbf{r} = c\, I_0 \left(\frac{\pi}{2}\right)^{3/2} \omega^2 \omega_z \\ \mathrm{Var}[I(t)] &= c \int_{\mathbb{R}^3} I^2(\mathbf{r})\mathrm{d}\mathbf{r} = c\, I_0^2 \left(\frac{\pi}{4}\right)^{3/2} \omega^2 \omega_z, \end{aligned} \qquad 14.$$

which can be derived by assuming that the particles are distributed in space according to a homogeneous Poisson point process. Thus, in the Gaussian limit, the concentration $c = (\pi^{3/2}\omega^2\omega_z)^{-1} \frac{E^2[I]}{\mathrm{var}[I]}$ and the brightness of a particle from the Fano factor $\mathrm{Var}[I]/E[I]$ can be measured.

FCS can also be used to obtain the reaction rate of a chemical process. Suppose we have a two-state reversible chemical reaction $A \rightleftharpoons B$, where A and B are the two states of the reaction. Let k_1^+ be the rate of A changing to B and k_1^- be the rate of B changing to A. This two-state reaction is typically described by a two-state continuous-time Markov chain where k_1^+ and k_1^- represent the (infinitesimal) transition rate. Suppose the two states A and B have different fluorescence intensities I_A and I_B. If we use X_t to denote the two-state process, then

$$\mathrm{Cov}[I(X_{t+\tau}), I(X_t)] = \mathrm{Var}[I(X_t)] \exp(-(k_1^+ + k_1^-)\tau).$$

This equation can be used to estimate the relaxation time $(k_1^+ + k_1^-)^{-1}$ of the reaction.

In the late 1980s, researchers started to measure non-Gaussian intensity distributions from small systems and obtain information about the heterogeneity of brightness in a mixture of particles. Various methods emerged, with acronyms such as FDS (fluorescence distribution spectroscopy), HMA (high-moment analysis), PCH (photon-counting histogram), and FIDA (fluorescence intensity distribution analysis). Non-Gaussian behavior means that higher-order temporal statistics such as $E[I(t_1 + t_2)I(t_1)I(0)]$ also contains useful information.

If $\Delta I(t) = I(t) - E[I]$ is a Markov process and is linear, i.e., the conditional expectation

$$E[\Delta I(t+\tau)|\Delta I(t) = z] = zg(\tau) \text{ with } g(0) = 1, \quad 15.$$

then the autocovariance function

$$E[\Delta I(t+\tau)\Delta I(t)] = E[(\Delta I)^2]g(\tau). \quad 16.$$

Therefore, we see that the functional form of the autocorrelation function (Equation 16) and the relaxation function after perturbation (Equation 15) are the same. This is the mathematical basis of the traditional, phenomenological approach of Einstein, Onsager, Lax, and Keizer to fluctuations. In a similar spirit, the higher-order temporal correlation functions are mathematically related to relaxations with multiple perturbations, known as multidimensional spectroscopy (Wiener 1966, Ridgeway et al. 2012).

The experimentally determined fluorescence autocorrelation function $\hat{g}(n\delta)$, where $n = 1, 2, \ldots$ and δ is the time step for successive measurements, often has a curious feature: The measured $\hat{g}(0)$ is always much greater than the extrapolated value from $\hat{g}(n)$ based on $n \geq 1$. In fact, the difference, known as shot noise, is approximately $E(I)$. Its origin is the Poisson nature of the random emissions of fluorescent photons, which are completely uncorrelated on the timescale of δ. Instead of treating the experimental fluorescence reading as a deterministic function of the underlying X_t, one needs to consider the quantum nature of photon emission—the photon counts are Poisson with the intensity function as the mean. Accordingly, the photon count from a single diffusing particle is an integer random variable with distribution (Qian 1990)

$$\Pr(I_1(t) = k) = \int_\Omega \frac{I^k(\mathbf{r})}{k!} e^{-I(\mathbf{r})} f_X(\mathbf{r}, t)\mathrm{d}\mathbf{r},$$

where $f_X(\mathbf{r}, t)$ is the probability density function of X_t. Therefore, under the assumption that Brownian particles are uniformly distributed in space,

$$E[I_1] = \frac{1}{|\Omega|}\int_\Omega I(\mathbf{r})\mathrm{d}\mathbf{r}, \quad \mathrm{Var}[I_1] = \frac{1}{|\Omega|}\int_\Omega (I(\mathbf{r}) + I^2(\mathbf{r}))\mathrm{d}\mathbf{r} - E^2[I_1].$$

Now again consider total N i.i.d. particles, and let $N, \Omega \to \infty$ and $N/|\Omega| = c$. Assuming that the particles are distributed in space according to a homogeneous Poisson point process, we have

$$E[I_1] = c\int_{\mathbb{R}^3} I(\mathbf{r})\mathrm{d}\mathbf{r}, \quad \mathrm{Var}[I_1] = E[I] + c\int_{\mathbb{R}^3} I^2(\mathbf{r})\mathrm{d}\mathbf{r}.$$

Comparing this with Equation 14, we see the extra shot-noise term $E[I]$. This is a good example of the textbook problem regarding the sum of a random number of independent random variables. In a laser-illuminated region, there is a random number of fluorescent particles, and each particle emits a Poisson number of photons; thus, the total photon count is a sum of a random number of terms.

Recently, the optical setup for FCS has been expanded to have two different colors of fluorescence or to have two laser beams at different locations of the system (Schwille et al. 1997, Dertinger et al. 2007). These measurements generate multivariate stationary fluorescence fluctuations. There are good opportunities for in-depth statistical studies of the new data, for example, the assessment of time reversibility of a Gaussian process (Qian 2001, Qian & Elson 2004).

5. DISCRETE MARKOV DESCRIPTION OF SINGLE-MOLECULE KINETICS

Diffusion theory describes a continuous-state, continuous-time Markov process (Wax 1954, Schuss 2010). In the 1970s, motivated mainly by novel experimental data from single-channel recordings

of membrane protein conductance, intense studies began to explore discrete-state continuous-time Markov processes [also called a Q-process by Doob (1942) and Reuter (1957)] as models for internal stochastic dynamics of individual biomacromolecules. For their contributions, E. Neher and B. Sakmann received the Nobel Prize in 1976. Sakmann & Neher (2009) provide a thorough review of single-channel recording. We also refer readers to earlier accounts of the development of a discrete-state Markov approach in biochemistry prior to the single-channel era (Bharucha-Reid 1960, McQuarrie 1967) and an exhaustive summary of the literature on ion-channel modeling and statistical analysis (Ball & Rice 1992).

Enzymes and proteins are large molecules consisting of tens of thousands of atoms, which are sometimes called biopolymers (see also Section 6). One of the central concepts established since the 1960s is that a protein can have several discrete conformational states: These states have different atomic arrangements within the molecule, and they can be observed through their various molecular characteristics, including absorption and emission optical spectra, physical sizes, or biochemical functional activities. These different probes can have different temporal resolutions and sensitivities. Using a highly sensitive probe with reasonably high temporal resolution, one can measure the dynamic fluctuations of a single protein as a stationary, discrete-state stochastic process. Therefore, Markov, or hidden Markov, models are natural tools for describing the conformational dynamics of a protein and such measurements.

5.1. Single-Channel Recording of Membrane Proteins

The earliest single-molecule experiments were carried out in the 1970s on ion channels; the patch-clamp technique pioneered by Neher and Sakmann enables reliable recording of membrane protein conductance on a single channel. As the closing and opening of an ion channel control the passage of ions across a cell membrane, the conductance recorded in experiments essentially consists of step functions, such as (stochastically) alternating high and low current levels. The simplest model to describe such on-off signaling is the two-state continuous-time Markov chain model

$$\text{open} \rightleftharpoons \text{close}. \qquad 17.$$

Owing to experimental noise and data filtering, the sequence of real observations $\{y(t_i), i = 1, 2, \ldots\}$ is better described by hidden Markov models. Under specific models, such as $y(t_i)|X(t_i) \sim N(X(t_i), \sigma^2)$, where $X(t)$ is the underlying state of the ion channel, the MLE for the transition rates can be obtained in straightforward fashion.

The conductance of real ion channels, however, is typically much more complicated than that given by the simple two-state model. For example, in addition to the open and closed states of the ion channel, there may exist blocked states, in which the binding of a blocking molecule to the ion channel stops the ion flow. Alternatively, the opening of a channel may be triggered by the binding of an agonist molecule. An ion channel, thus, could have multiple closed and open states. A complication for modeling and inference is that these open (and closed) states are not distinguishable from experimental data: Typically, the open (and closed) states have the same conductance. Therefore, we are dealing with aggregated Markov processes: Although the underlying mechanism is Markovian, we observe only in which aggregate (i.e., collection of states) the process is found (Fredkin & Rice 1986). A natural issue regards the identifiability of different models, given that we can observe only the aggregates. Note, that it is quite possible that two distinct models may give the same data structure/likelihood.

Statistical questions include how to estimate the number of (open and closed) states, postulate a model, and infer the parameters of the model. Ball & Rice (1992) surveyed the statistical analysis and modeling of the ion channel data. Chapter 3 and part III of the encyclopedic book by Sakmann

& Neher (2009) provide an introduction and review of ion channel data analysis, from initial data processing to inference complications, such as the time interval omission problem.

Parallel to the construction, testing, and estimation of Markov models, an alternative statistical approach is to treat the inference as a change-point detection problem: Given the on-off signal, use the data to determine the change points (i.e., the transition times) and then infer the sojourn times and their correlation, which provide clues for eventual model building. The change-point approach can be viewed as nonparametric because it does not explicitly rely on a (Markov) model specification. The problem of change-point estimation has a long history in statistics dating back to the 1960s. More recent approaches particularly relevant for single-channel data include use of the BIC (Bayesian information criterion) penalty (Yao 1988), the quasi-likelihood method (Braun et al. 2000), the L_1 penalty method (Tibshirani et al. 2005), the multiresolution method (Hotz et al. 2012), and the marginal likelihood method (Du et al. 2013). Compared with the parametric inference methods based on continuous-time Markov chains, many of these change-point methods are flexible and can be made automatic. Thus, they are suitable for fast initial analysis of a large amount of single-channel data, such as the thousands of data traces commonly generated in modern single-channel recording experiments.

5.2. Two-State and Three-State Single-Molecule Kinetics

A two-state Markov chain, such as in Equation 17, is widely used in biochemical kinetics. It is typically diagrammed as

$$A \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} B, \tag{18.}$$

where A and B are the two states and k_1^+ and k_1^- are the (infinitesimal) transition rates. One of the simplest biochemical reactions is the reversible binding of a single protein E to its substrate molecule S, $E + S \rightleftharpoons ES$. It can often be described by this two-state Markov model with rate parameters $k_1^+ = k_1^0 c_S$ and k_1^-, where c_S denotes the concentration of the substrate molecules. In this case, $k_1^+ = k_1^0 c_S$ assumes that the protein concentration is sufficiently dilute. Thus, a large number of substrate molecules S per E must be present for the concentration c_S to remain essentially constant. Writing out $k_1^+ = k_1^0 c_S$ also highlights the fact that the concentration c_S of the substrate can be controlled in experiments. Thus, one can study the effect of the concentration c_S on the overall reaction. In chemical kinetics, k_1^0 and k_1^+ are called second-order and pseudo-first-order rate constants, respectively: A second-order rate constant has the dimension $[\text{time}]^{-1} \times [\text{concentration}]^{-1}$, whereas a first-order rate constant has the dimension $[\text{time}]^{-1}$. The states E and ES of a single protein can be monitored through a change in the fluorescence intensity of the molecule, for example, through either the intrinsic fluorescence of the protein or the Föster resonance energy transfer between the protein and the substrate.

A three-state Markov chain is often used to describe an enzyme's cycling through three states, E, ES, and EP:

$$E + S \underset{k_1^-}{\overset{k_1^0}{\rightleftharpoons}} ES, \quad ES \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} EP, \quad \text{and } EP \underset{k_3^0}{\overset{k_3^+}{\rightleftharpoons}} E + P. \tag{19.}$$

An enzyme catalytic cycle is completed every time it helps convert a substrate molecule S to a product P, while the state of the enzyme molecule returns to the E so that it can start the cycle to convert the next substrate molecule (shown in **Figure 1**). The enzyme E serves as a catalyst to the chemical

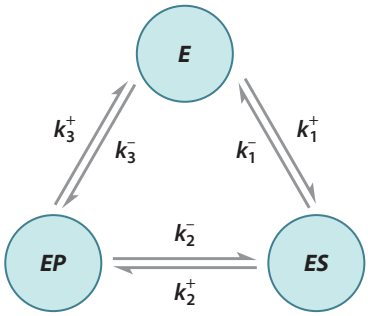

Figure 1

Typical enzyme kinetics can be written as a sequence of biochemical steps, as in Equation 19, or from a single-enzyme perspective, it can be expressed as a cycle, as illustrated here. Note that the second-order rate constants k_1^0 and k_3^0 in Equation 19 have been replaced by pseudo-first-order rate constants k_1^+ and k_3^-, respectively. The simplest statistical kinetics model is to consider this system as a continuous-time, discrete-state Markov process. A more sophisticated model, when there are sufficient data, could be a semi-Markov model with arbitrary, nonexponential sojourn time for each of the three states (Wang & Qian 2007).

transformations $S \rightleftharpoons P$. Again, using the idea of pseudo-first-order rate constants, we have the (infinitesimal) transition rates $k_1^+ = k_1^0 c_S$ and $k_3^- = k_3^0 c_P$, where c_P is the concentration of the product P.

A three-state Markov process is reversible if $k_1^+ k_2^+ k_3^+ / (k_1^- k_2^- k_3^-) = 1$, which is a special case of the Kolmogorov criterion of reversibility (Kelly 1979). This mathematical concept precisely matches the important notion of a chemical equilibrium between S and P when

$$\left(\frac{c_P}{c_S}\right)^{eq} = \frac{k_1^0 k_2^+ k_3^+}{k_1^- k_2^- k_3^0}. \qquad 20.$$

In fact, as is widely known in biochemistry, reaction $S \leftrightarrows P$ will have very small forward and backward first-order rate constants α^+ and α^- when the enzyme is absent. Nevertheless, the fundamental law of chemical equilibrium dictates that $\alpha^+/\alpha^- = k_1^0 k_2^+ k_3^+ / (k_1^- k_2^- k_3^0)$ (Lewis 1925).

In a living cell, however, the substrate and the product of an enzyme are usually not at their chemical equilibrium, and their concentrations c_S and c_P do not satisfy the equality in Equation 20. Accordingly,

$$\frac{k_1^+ k_2^+ k_3^+}{k_1^- k_2^- k_3^-} = \frac{k_1^0 c_S k_2^+ k_3^+}{k_1^- k_2^- k_3^0 c_P} \neq 1,$$

for which the corresponding Markov chain is no longer reversible. This finding motivated the mathematical theory of nonequilibrium steady state (NESS) (Jiang et al. 2004, Ge et al. 2012, Zhang et al. 2012). For a strongly irreversible, three-state Markov process, its Q-matrix (i.e., the infinitesimal generator) may have a pair of complex eigenvalues, giving rise to a nonmonotonic, oscillatory autocorrelation function (Qian & Elson 2002). For example, if $k_1^- = k_2^- = k_3^- = 0$ and $k_1^+ = k_2^+ = k_3^+ = 1$, then the two nonzero eigenvalues are $-\frac{1}{2}(3 \pm i\sqrt{3})$. Such oscillatory behavior has been observed in single-molecule experiments.

5.3. Entropy Production and Nonequilibrium Steady State

The chemical NESS also motivated the mathematical concept of entropy production rate (M.-P. Qian et al. 1991, Jiang et al. 2004):

$$e_p = \lim_{t\to\infty} \frac{1}{t} \ln\left(\frac{d\mathbb{P}_t}{d\mathbb{P}_t^-}\right). \qquad 21.$$

For a continuous-time Markov process $X(t)$, $\mathbb{P}_t$ in Equation 21 is the likelihood of a stationary trajectory, and $\mathbb{P}_t^-$ is the likelihood of the time-reversed trajectory. For example, if $\mathbb{P}_t$ is the likelihood of a particular trajectory $2 \to 3 \to 1$, where the transitions occur at t_1 and t_2 with $0 < t_1 < t_2 < t$, then $\mathbb{P}_t^-$ is the likelihood of the trajectory $1 \to 3 \to 2$, where the transitions occur at $t - t_2$ and $t - t_1$.

For a three-state system, it is easy to show that

$$e_p = J^{\text{NESS}} \ln\left(\frac{k_1^+ k_2^+ k_3^+}{k_1^- k_2^- k_3^-}\right), \tag{22.}$$

with NESS probability circulation

$$J^{\text{NESS}} = \frac{k_1^+ k_2^+ k_3^+ - k_1^- k_2^- k_3^-}{\left\{\begin{array}{c} k_1^+ k_2^+ + k_1^- k_3^- + k_2^+ k_3^- + k_2^+ k_3^+ + k_2^- k_1^- \\ + k_3^+ k_1^- + k_3^+ k_1^+ + k_3^- k_2^- + k_1^+ k_2^- \end{array}\right\}}.$$

We see that e_p is never negative, and it is zero if and only if the Markov process is reversible. In fact, in the energy unit of $k_B T$, the logarithmic term in Equation 22 is the chemical potential difference between S and P: $\Delta\mu_{S\to P} = k_B T \ln \frac{k_1^+ k_2^+ k_3^+}{k_1^- k_2^- k_3^-}$, J^{NESS} is the number of reactions per unit time, and e_p is the amount of heat dissipated into the environment per unit time. The chemical potential equaling heat dissipation is the first law of thermodynamics; $e_p \geq 0$ is interpreted as the second law of thermodynamics. The second law has always been taught as an inequality; Equation 21 provides a more quantitative formulation in terms of a Markov process. For finite t, e_p in Equation 21 is stochastic and has a negative tail. Characterizing this negative tail under a proper choice of the initial probability for a finite trajectory is the central theme of the recently developed fluctuation theorems (Kim 2011, Seifert 2012).

5.4. Michaelis–Menten Single-Enzyme Kinetics

In single-molecule enzyme kinetics (Lu et al. 1998), one can measure the arrival times of successive product P, following the simple Michaelis–Menten (MM) enzyme kinetic scheme (Kou et al. 2005, English et al. 2006):

$$E + S \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} ES, \quad ES \underset{k_3^-}{\overset{k_3^+}{\rightleftharpoons}} E + P. \tag{23.}$$

This is a simpler model than that in Equation 19: It is assumed that reactions associated with k_2^+ and k_2^- are so fast that they can be neglected. Because each arriving P is immediately processed, $k_3^- = k_3^0 c_P = 0$. The arrivals of every P are now a renewal process with mean waiting time $E[T]$ easily computed (Qian & Elson 2002, Kou et al. 2005, Qian 2008) from

$$E[T] = \frac{1}{k_1^+} + \frac{1}{k_1^- + k_3^+} + \frac{k_1^-}{k_1^- + k_3^+} E[T] + \frac{k_3^+}{k_1^- + k_3^+} 0.$$

Solving $E[T]$ and noting $k_1^+ = k_1^0 c_S$, one obtains

$$E^{-1}[T] = \frac{V_{\max} c_S}{K_M + c_S}, \quad V_{\max} = k_3^+, \quad K_M = \frac{k_1^- + k_3^+}{k_1^0}. \tag{24.}$$

This is the celebrated MM equation for steady-state enzyme catalytic velocity, first discovered in 1913 on the basis of a nonstatistical theory. One of the immediate insights from the probabilistic derivation of the MM equation is that, if an enzyme has only a single unbound state E, then irrespective of how many and how complex the bounding states $(ES)_1, \ldots, (ES)_n$ may be, the MM

equation is always valid. The expressions for the $V_{\max}$ and K_M can be very complex (English et al. 2006, Min et al. 2006). We discuss in some detail the single-molecule experiments on enzymes and models beyond the MM mechanism in Section 5.5. If $c_P \neq 0$, then the NESS probability circulation in the enzyme cycle is as follows (Qian 2008):

$$J^{\text{NESS}} = \frac{(V_{\max}/K_M)c_S - (V^-_{\max}/K^P_M)c_P}{1 + c_S/K_M + c_P/K^P_M}, \quad V^-_{\max} = k^-_1, \quad K^P_M = \frac{k^-_1 + k^+_3}{k^0_3}.$$

This is known as the Briggs–Haldane equation for a reversible enzyme.

5.5. Single-Molecule Enzymology in Aqueous Solution

We have seen how the schemes noted in Equations 19 and 23 describe enzyme kinetics. Traditionally, they are used to set up (coupled) differential equations, which specify how the concentrations of the enzyme, the substrate, and the product change over time. These theoretical descriptions then can be compared with experimental results carried out in bulk solution, which involve a large ensemble of enzyme molecules.

In contrast to these traditional ensemble experiments, observing the action of a single enzyme molecule in aqueous solution requires the development of methods that immobilize an enzyme molecule and make the experimental system fluorescent as well as the use of high-sensitivity optical microscopy. This was first accomplished in 1998 (Lu et al. 1998) on cholesterol oxidase, where the active site of the enzyme, $E + S$ and ES in Equation 23, is fluorescent, yielding an on-off system. The experimental data of Lu et al. (1998) have an appearance similar to that of the on-off data from ion channels (Section 5.1). Thus, many data analysis tools developed for single-channel recording can be applied. Experimental fluorescence techniques, such as the design and utilization of a fluorescent substrate, fluorescent active site, and fluorescent product, and experimental techniques to immobilize an enzyme molecule were reviewed in Xie & Lu (1999) and Xie (2001), which also discuss the relationship between single-molecule enzymology and the traditional ensemble approach.

As experimental methods have developed and matured, we can finally study and test the MM mechanism directly (Equation 23) on the single-molecule scale. Using a fluorescent product, English et al. (2006) conducted single-molecule experiments on the enzyme β-galactosidase. The sharp fluorescence spikes from the product enable experimental resolution of β-galactosidase's individual turnovers (i.e., the successive cycles of the enzyme). Experimental data showed that (*a*) the distribution of the enzyme's turnover times is much heavier than an exponential distribution, contradicting the MM mechanism's prediction; (*b*) there is a strong serial correlation in a single enzyme's successive turnover times, also contradicting the MM mechanism; and (*c*) the hyperbolic MM relationship of $E^{-1}(T) \propto c_S/(c_S + K_M)$, as given in Equation 24, still holds. To explain these experimental results, particularly their contradiction with the MM mechanism, Kou et al. (2005) introduced the following model (diagrammed in **Figure 2**).

In **Figure 2**, $E_1, E_2, \ldots$ represent the different conformations of the enzyme, and SE_i are the different conformations of the enzyme-substrate complex. The model is based on the insight that a protein molecule can have multiple conformational states: These states have different atomic arrangements and can have different biochemical functional activities. Detailed calculations (see Kou et al. 2005, English et al. 2006, Kou 2008b, Du & Kou 2012) show that the model is capable of explaining the experimental data.

Data from experiments such as those of English et al. (2006) have different patterns from the on-off data of Lu et al. (1998). The fluorescent product, which, once formed, quickly diffuses away from the focus of the microscope, is used in these experiments. Thus, the experimental

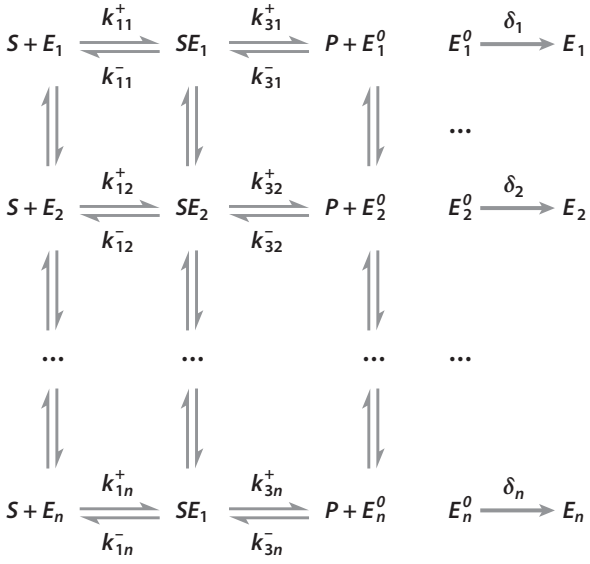

Figure 2

A discrete schematic illustrating the Markovian kinetics of a single enzyme molecule with conformational fluctuations.

data consist of fluorescent spikes, with each spike corresponding to the formation of one product molecule, amid fluorescence from the background. In principle, the time lag between two successive spikes is the (individual) turnover time of the enzyme. In practice, because the level of the fluorescent spike is random (as a product molecule spends a random time in the focal area of the microscope before diffusing away), one needs to threshold the data to locate the spikes. Finding a statistically efficient threshold level (to minimize false positives) for such data is an open problem.

5.6. Motor Proteins with Mechanical Movements Against External Force

Particular enzymes called motor proteins can move along their designated linear, periodic tracks inside a living cell, even against a resistant force. The energy of the motor is derived from the chemical potential in the $S \to P$ reaction, given in Equation 22 (Qian 1997, 2005; Fisher & Kolomeisky 1999; Kolomeisky & Fisher 2007; Chowdhury 2013). An external mechanical force $F_{\rm ext}$ enters the rate constants to effect the conformational transition of a motor protein as follows: If the transition from conformational state A to state B moves a distance d_{AB} along the track against the force, then according to Boltzmann's law

$$\frac{k_{A\to B}(F_{\rm ext})}{k_{B\to A}(F_{\rm ext})} = \frac{k_{A\to B}(0)}{k_{B\to A}(0)} \exp\left(-\frac{F_{\rm ext} d_{AB}}{k_{\rm B}T}\right).$$

Substitute such a relation into Equation 22 and let d be the total motor step length for one enzyme cycle (from S to P), then

$$e_p = J^{\rm NESS} \times \frac{1}{k_{\rm B}T}\left(\Delta\mu_{S\to P} - F_{\rm ext} d\right).$$

In this case, part of the chemical energy from transformation $S \to P$ is converted to mechanical energy. The part that becomes heat is the entropy production.

The motor protein carries out a biased random walk with velocity $v_{\text{motor}} = J^{\text{NESS}} d$. With increasing force F_{ext}, v_{motor} decreases. When $F_{\text{ext}} = \Delta\mu_{S\to P}/d$, the random walk is no longer biased; this is known as a stalling force. One can also compute the dispersion of the motor, i.e., a diffusion coefficient:

$$D_{\text{motor}} = \frac{d^2}{2E[T_c]}, \quad E[T_c] = \frac{1 + c_S/K_M + c_P/K_M^P}{(V_{\max}/K_M)c_S + (V_{\max}^-/K_M^P)c_P}.$$

In fact, as a semi-Markov process (also known as Markov renewal process or continuous-time random walk), the mean cycle time is $E[T_c]$ and the ratio of probabilities of forward and backward cycles is $\frac{(V_{\max}/K_M)c_S}{(V_{\max}^-/K_M^P)c_P}$.

5.7. Advanced Topics

This brief section highlights some more advanced statistical topics from single-molecule analysis. Most of these works have been described in the physics and physical chemistry literature; the analyses remain to be properly treated in the hands of statisticians.

5.7.1. Empirical measure with finite time. Even for the simplest two-state Markov process, some of the statistics can be complex. For example, Geva & Skinner (1998) analytically studied the statistical quantity

$$X_\tau = \frac{1}{\tau}\int_0^\tau \xi_B(t)\,\mathrm{d}t,$$

where $\xi_B(t)$ is the indicator function for state B in Equation 18. They showed that the pdf (probability density function) of X_τ can be obtained in terms of its Fourier transform $\gamma(y)$:

$$\gamma(y) = e^{-\frac{1}{2}(k\tau+iy)}\left[\cosh\,\phi + \left(\frac{\alpha}{\phi}\right)\sinh\phi\right],$$

where $k = k_1^+ + k_1^-$, $\alpha = \frac{1}{2}k\tau - i\left(p - \frac{1}{2}\right)y$, $p = k_1^+/(k_1^+ + k_1^-)$, and

$$\phi^2 = \left(\frac{k\tau}{2}\right)^2 - i\left(p - \frac{1}{2}\right) - \left(\frac{y}{2}\right)^2.$$

Thus, for large τ,

$$\gamma(y) = e^{-\frac{\sigma^2(\tau)y^2}{2} - ipy}, \quad \sigma^2(\tau) = \frac{2p(1-p)}{k\tau}.$$

5.7.2. Non-Markovian two-state systems. Some enzymes exhibit clear two-state stochastic behavior, but the process is not Markovian. For example, the consecutive dwell times in state B could have nonzero correlation (Lu et al. 1998). This is a strong violation of the Markovian property. To explain this observation, the theory of dynamic disorder, or fluctuating enzyme, assumes that k_1^+ and k_1^- in Equation 18 are stochastic processes in the form $k_1^\pm(t) = \tilde{k}_1^\pm e^{-X_t}$, where X_t is an Ornstein-Uhlenbeck process (see Equation 3) (Agmon & Hopfield 1983, Schenter et al. 1999, Kou et al. 2005). In this case, even though $\xi_B(t)$ is no longer a Markov process, (ξ_B, X) together is now a coupled diffusion process (Qian 2002). A more complex model on X_t (describing it as fractional Gaussian noise) is considered in Wang & Wolynes (1995). One can also model X_t by the GLE (Kou & Xie 2004) of Section 2.2.

5.7.3. Dwell time distribution peaking. As discussed above, a continuous-time Markov chain in an NESS can have complex eigenvalues. Thus, the power spectrum of its stationary data

can exhibit an off-zero peak representing intrinsic frequency (Qian & Qian 2000). However, a surprising result discovered independently by Li & Qian (2002) and Tu (2008) is that one can also observe an off-zero peak in the pdf of the dwell time within a group of states, which is impossible for a reversible process.

5.7.4. Detailed balance violation and event ordering. The fundamental insight that a sustained chemical energy input is necessary to observe an irreversible Markov process in molecular systems has opened several lines of inquiry on stationary data. On the one hand, for stationary molecular fluctuations in chemicothermodynamic equilibrium, one wants to test the preservation of detailed balance (Rothberg & Magleby 2001, Witkoskie & Cao 2006, Nagy & Tóth 2012). On the other hand, for a molecular process with unknown mechanism, one wants to discover whether it is chemically driven (Qian & Elson 2004). In fact, quantification of the deviation from reversibility could reveal the source of the external energy supply. Finally, for a system with breakdown of detailed balance, event ordering from statistical analysis provides insights regarding the molecular mechanism (Sisan et al. 2010).

The concept of detailed balance also exists in chemistry (Fowler & Milne 1925, Lewis 1925, Feinberg 1989), but it is essentially different from the same term known in statistics. For a chemical detailed balance, a set of linear and nonlinear reactions forming a reaction cycle must have zero cycle flux in chemical equilibrium. This chemical detailed balance is expressed in terms of the concentrations of the reactants, which are deterministic quantities. There is no probability involved in this statement. If all the reactions are unimolecular, however, then a chemical reaction system in terms of the law of mass action is equivalent to a continuous-time Markov chain. Only in this case are the chemical and the probabilistic detailed balance conditions the same.

6. POLYMER DYNAMICS AND GAUSSIAN PROCESSES

Polymer dynamics is another highly successful theory based on stochastic processes (Flory 1969, Doi & Edwards 1988). A polymer chain in aqueous solution is modelled by a string of identical beads connected by harmonic springs. The Langevin equation for the k^{th} bead ($k = 1, 2, \ldots, N$) is

$$m\frac{\mathrm{d}^2 X_k}{\mathrm{d}t^2} + \zeta\frac{\mathrm{d}X_k}{\mathrm{d}t} = \alpha\,(X_{k-1} - 2X_k + X_{k+1}) + \sqrt{2\zeta k_{\mathrm{B}}T}\,\frac{\mathrm{d}B_k(t)}{\mathrm{d}t}, \qquad 25.$$

where α is the spring constant, m and ζ are the mass and damping coefficient of a bead, and $B_k(t)$ are i.i.d. Wiener processes, again representing the collisions with the solvent. Usually, the mechanical system is under an overdamped condition, e.g., $m\alpha \ll \zeta^2$, in which the acceleration is negligible. Then Equation 25 is simplified to

$$\zeta\frac{\mathrm{d}X_k}{\mathrm{d}t} = \alpha\,(X_{k-1} - 2X_k + X_{k+1}) + \sqrt{2\zeta k_{\mathrm{B}}T}\,\frac{\mathrm{d}B_k(t)}{\mathrm{d}t}, \qquad 26.$$

which is a multidimensional Ornstein–Uhlenbeck process. A polymer molecule represented by such a dynamical model is called a Gaussian chain.

The boundary condition $X_0(t) = 0$ is used to represent a tethered polymer end, and $X_N(t) = X_{N+1}(t)$ represents a free polymer end. To study Equation 26, an elegant approach is to approximate it using a stochastic partial differential equation:

$$\zeta\frac{\partial X(s,t)}{\partial t} = \alpha\frac{\partial^2 X(s,t)}{\partial s^2} + \sqrt{2\zeta k_B T}\,\frac{\mathrm{d}B(s,t)}{\mathrm{d}t},$$

where $\frac{\mathrm{d}}{\mathrm{d}t}B(s,t)$ represents spatiotemporal white noise. With the boundary conditions $X(0,t)=0$ and $\frac{\partial X(L,t)}{\partial s}=0$, the Fourier transform yields

$$X(s,t)=\sum_{j=0}^{\infty}\xi_j(t)\sin(\lambda_j s),\ \lambda_j=\left(j+\frac{1}{2}\right)\frac{\pi}{L},$$

where each normal mode

$$\zeta\frac{\mathrm{d}\xi_j(t)}{\mathrm{d}t}=-\alpha\lambda_j^2\xi_j(t)+F_j(t),$$

and

$$E\left[F_i(t)F_j(\tau)\right]=\left(\frac{4\zeta k_{\mathrm{B}}T}{L}\right)\delta_{ij}\delta(t-\tau).$$

Each $\xi_j(t)$ is an Ornstein–Uhlenbeck process; its stationary distribution has variance

$$\sigma_j^2=\frac{2k_{\mathrm{B}}T}{\alpha\lambda_j^2L}.$$

Therefore, $X(s,t)$ is a Gaussian random field with stationary variance

$$\sigma^2(s)=\sum_{j=0}^{\infty}\left(\frac{2k_{\mathrm{B}}T}{\alpha\lambda_j^2L}\right)\sin^2(\lambda_j s).$$

One strong prediction of the Gaussian polymer theory is that the end-to-end distance of a long polymer should be scaled as the square root of its molecular weight M. This result has become the standard against which a real polymer is classified: When a polymer is dissolved in a bad solvent, its conformation is more collapsed. Thus, its end-to-end distance may scale as M^{ν} with $\nu<1/2$. By contrast, owing to physical exclusion among polymer segments, a real polymer in a good solvent is expected to be more expanded with $\nu>1/2$. Indeed, the problem of excluded-volume effect in polymer theory has been a major topic in chemistry and mathematics. Paul Flory received the 1974 Nobel Prize in Chemistry for his studies leading to $\nu=3/5$. The rigorous mathematical work on this subject, known as self-avoiding random walks, was carried out by Wendelin Werner, who received the 2006 Fields Medal for related work.

6.1. Tethered Particle Motion Measuring DNA Looping

Polymer theory has been widely applied in the modeling of biomacromolecules, especially DNA (Schellman 1980). In the 1990s, Gelles, Sheetz, and their colleagues developed tethered particle motion (TPM), a single-molecule method to study transcription and DNA looping (Schafer et al. 1991, Finzi & Gelles 1995). In TPM, the trajectory of a Brownian motion particle is attached to a piece of DNA and followed. The statistical movements of the particle provide information on DNA flexibility, length, etc. The theory for TPM requires a boundary condition at X_N that is different from Equation 26, taking into account the much larger particle that serves as the optical marker (Qian & Elson 1999, Qian 2000).

6.2. Rubber Elasticity and Entropic Force

Gaussian chain theory owes its great success to the central limit theorem. The end-to-end distance of a polymer chain can be thought of as a sum of N i.i.d. random segments $\mathbf{l}_k$, $1\le k\le N$, where N is proportional to the total molecular weight M. As long as $\mathbf{l}$ has a distribution with finite second

moment, then

$$E\left[\left\|\sum_{k=1}^{N}\mathbf{l}_k\right\|^2\right]=\sum_{j=1}^{N}\sum_{k=1}^{N}E[\mathbf{l}_j\cdot\mathbf{l}_k]=N\sigma^2$$

(Flory 1969), where it is assumed that $E[\mathbf{l}_j\cdot\mathbf{l}_k]=\sigma^2\delta_{jk}$, owing to spatial symmetry.

We would like to point out that, to a large extent, the elasticity of rubber is not due to any other molecular interaction, but instead is simply a consequence of this statistical behavior of a Gaussian chain. The end-to-end distance is asymptotically a Gaussian random variable with variance $N\sigma^2$:

$$\frac{1}{\sqrt{2\pi N\sigma^2}}e^{-x^2/(2N\sigma^2)}\quad(\text{for large } N).$$

Let one end of a chain be attached. Then the stochastic chain dynamics, on average, pulls the free end from a less-probable position toward a more-probable one: This is called entropic force in polymer physics. In fact, reversing Boltzmann's law, there is an equivalent harmonic entropy potential energy $U(x)=k_{\mathrm{B}}T\,x^2/(2N\sigma^2)$ with spring constant $k_{\mathrm{B}}T/(N\sigma^2)$.

6.3. Potential of Mean Force and Conditional Probability

Stationary probability giving rise to an equivalent force is one of the fundamental insights from polymer chemistry. A key concept in statistical chemistry, first developed by Kirkwood (1935), is the potential of mean force. In essence, it is an incarnation of the conditional probability.

To illustrate the idea, let us again consider the Langevin equation for an overdamped particle in a potential $U(x)$:

$$\mathrm{d}X(t)=\frac{1}{\zeta}\left(-U'(X)\mathrm{d}t+\sqrt{2\zeta k_{\mathrm{B}}T}\,\mathrm{d}B(t)\right).$$

The corresponding Kolmogorov forward equation for the probability density function $f_X(x,t)$ is

$$\frac{\partial f_X(x,t)}{\partial t}=\frac{1}{\zeta}\frac{\partial}{\partial x}\left(k_{\mathrm{B}}T\,\frac{\partial f_X(x,t)}{\partial x}+\frac{\mathrm{d}U(x)}{\mathrm{d}x}f_X(x,t)\right),\qquad 27.$$

where $-U'(x)$ represents a potential force acting on the Brownian particle.

Now let us consider a Brownian particle in a three-dimensional space without any force. If one is interested only in the distance of the Brownian particle to the origin, $R(t)$, then the pdf $f_R(r,t)$ follows a Kolmogorov forward equation:

$$\frac{\partial f_R(r,t)}{\partial t}=\frac{k_{\mathrm{B}}T}{\zeta}\frac{\partial}{\partial r}\left(\frac{\partial f_R(r,t)}{\partial r}-\frac{2}{r}f_R(r,t)\right).\qquad 28.$$

Comparing Equation 28 with Equation 27, we see that the stochastic motion of $R(t)$ experiences an equivalent force $2k_{\mathrm{B}}T/r$, with a potential function $U_R(r)=-2k_{\mathrm{B}}T\ln r$. This is again an entropic force, and the corresponding $U_R(r)$ is called the potential of mean force. The entropic force arises essentially from a change of measure; therefore, it is fundamentally rooted in the theory of probability. The potential of mean force $U_R(r)$ should be understood as

$$U_R(r)=-k_{\mathrm{B}}T\,\ln\{\text{conditional stationary prob. given } R=r\}+\text{const.}\qquad 29.$$

Equation 29 is again applying the Boltzmann's law in reverse, relating an energy to probability.

7. STATISTICAL DESCRIPTION OF GENERAL STOCHASTIC DYNAMICS

7.1. Chemical Kinetic Systems as a Paradigm for Complex Dynamics

It is arguable that, since the work of Kramers, chemists have been among the first groups to fully appreciate the nature of separation of timescales in complex dynamics: Whereas the rapid atomic movements in a molecule are extremely fast, on the order of pico- to femtoseconds, a chemical reaction that involves passage through a saddle point in the energy landscape on this timescale is a rare event. From this realization, the notions of transition state and reaction coordinate have become two of the most elusive, yet extremely important, distinctly chemical concepts. Yet, they are even more important in biophysics, which, among other disciplines, deals with the transitions between conformational states of proteins. Although not widely articulated, this is the appropriate statistical treatment of any dynamic system with a separation of timescales due to statistical multimodality.

7.2. General Markov Dynamics with Irreversible Thermodynamics

Ever since the work of Kolmogorov, reversible, or symmetric, Markov processes have been widely studied both in theory and in applications. Detailed balance is one of the most important concepts in the theory of Markov chain Monte Carlo (MCMC). By contrast, the notion of entropy has grown increasingly prominent in general discussions on complex systems, usually in connection to information theory.

The central role of irreversible Markov processes in describing complex biophysical processes is now firmly established. In recent years, it has also become clear that entropy and entropy production are essential concepts in irreversible, often stationary, Markov processes. In this section, we give a concise description of this emergent statistical dynamic theory. We present only key results and leave out all mathematical proofs, which can be found elsewhere (see Qian et al. 2002, Esposito & van den Broeck 2010, Ge & Qian 2010, Qian 2013a).

Consider a diffusion process with its Kolmogorov forward equation in the form of

$$\frac{\partial f(\mathbf{x},t)}{\partial t} = \nabla\cdot(D(\mathbf{x})\nabla f(\mathbf{x},t) - b(\mathbf{x})f(\mathbf{x},t)) = \mathcal{L}[f]. \qquad 30.$$

Assume that it has an ergodic, differentiable stationary density $f^{\rm NESS}(\mathbf{x})$, $x \in \Omega$. Then one can define two essential thermodynamic quantities: the internal energy of the system $U(\mathbf{x}) = -\ln f^{\rm NESS}(\mathbf{x})$ and the entropy of the entire system

$$S[f(\mathbf{x},t)] = -\int_\Omega f(\mathbf{x},t)\ln f(\mathbf{x},t)\mathrm{d}\mathbf{x}.$$

Given the expected value of the U and the so-called generalized free energy $\Psi[f(\mathbf{x},t)] = E[U]-S$,

$$E[U](t) = \int_\Omega U(\mathbf{x})f(\mathbf{x},t)\mathrm{d}\mathbf{x}, \quad \Psi[f(\mathbf{x},t)] = \int_\Omega f(\mathbf{x},t)\ln\left(\frac{f(\mathbf{x},t)}{f^{\rm NESS}(\mathbf{x})}\right)\mathrm{d}\mathbf{x}. \qquad 31.$$

As relative entropy, the importance of $\Psi \geq 0$ is widely known, yielding the following set of equations that constitute a theory of irreversible thermodynamics:

$$\frac{\mathrm{d}\Psi}{\mathrm{d}t} = E_{\rm in} - e_p \leq 0, \quad \frac{\mathrm{d}S}{\mathrm{d}t} = e_p - b_{\rm ex}, \quad E_{\rm in}, e_p \geq 0; \qquad 32a.$$

$$E_{\rm in}(t) = \int_\Omega (\nabla \ln f^{\rm NESS}(\mathbf{x}) - D^{-1}(\mathbf{x})\mathbf{b}(\mathbf{x}))\mathbf{J}(\mathbf{x},t)\mathrm{d}\mathbf{x}; \qquad 32b.$$

$$e_p(t) = \int_\Omega (\nabla \ln f(\mathbf{x}, t) - D^{-1}(\mathbf{x})\mathbf{b}(\mathbf{x}))\mathbf{J}(\mathbf{x}, t)\mathrm{d}\mathbf{x}; \tag{32c.}$$

$$h_{\mathrm{ex}}(t) = \int_\Omega \mathbf{b}(\mathbf{x})D^{-1}(\mathbf{x})\mathbf{J}(\mathbf{x}, t)\mathrm{d}\mathbf{x}; \tag{32d.}$$

and

$$\mathbf{J}(\mathbf{x}, t) = \mathbf{b}(\mathbf{x})f(\mathbf{x}, t) - D(\mathbf{x})\nabla f(\mathbf{x}, t). \tag{32e.}$$

The first equation in Equation 32a can be interpreted as an energy balance equation, with the non-negative E_{in} and e_p as a source and a sink, respectively. e_p is called entropy production. The second equation in Equation 32a is an entropy balance equation, with heat exchange h_{ex} that can be either positive or negative. $\mathrm{d}\Psi/\mathrm{d}t \leq 0$ is the second law of thermodynamics.

For a reversible Markov process, $E_{\mathrm{in}}(t) \equiv 0$ for all t. Its stationary version has $\mathcal{J}(x) \equiv 0$ for all x and $e_p = h_{\mathrm{ex}} = 0$. This is known as chemicothermodynamic equilibrium in biophysics. In general, in an NESS, $\nabla \cdot \mathbf{J}^{\mathrm{NESS}} = 0$, but $\mathbf{J}^{\mathrm{NESS}} \neq 0$.

We now turn our attention to the dynamic Equation 30. Its generator is $\mathcal{L}^* = \nabla \cdot D(\mathbf{x})\nabla + \mathbf{b}(\mathbf{x})\nabla$. Introducing the inner product

$$(\phi, \psi) = \int_\Omega \phi(\mathbf{x})\psi(\mathbf{x})f^{\mathrm{NESS}}(\mathbf{x})\mathrm{d}\mathbf{x},$$

we find that the linear differential operator $\mathcal{L}^*$ can be decomposed into $\mathcal{L}^* = \mathcal{L}_s^* + \mathcal{L}_a^*$, a symmetric and an antisymmetric part, respectively. Thus, one has the operator in Equation 30, $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_a$:

$$\mathcal{L}_s[u] = \nabla \cdot (D(\mathbf{x})\nabla u(\mathbf{x}) - (D(\mathbf{x})\nabla \ln f^{\mathrm{NESS}}(\mathbf{x}))u(\mathbf{x})), \tag{33a.}$$

$$\mathcal{L}_a[u] = \nabla \cdot ((D(\mathbf{x})\nabla \ln f^{\mathrm{NESS}}(\mathbf{x}) - \mathbf{b}(\mathbf{x}))u(\mathbf{x})). \tag{33b.}$$

In connection to the thermodynamics in Equation 32, a diffusion process with pure $\mathcal{L}_s$ has $E_{\mathrm{in}}(t) = 0$; a process with pure $\mathcal{L}_a$ has $\mathrm{d}\Psi/\mathrm{d}t = 0$ for all t. Note that the operator in Equation 33b is hyperbolic rather than elliptical: It is a generalization of conservative, classical Hamiltonian dynamics (Qian 2013a). Equation 33a is a generalization of the heat kernel. The generalized Markov dynamics, therefore, unifies the Newtonian conservative and Fourier's dissipative dynamics.

Thermodynamics and the notions of dissipative and conservative dynamics have been the cornerstones of classical physics. We now see that they emerged from a statistical description of Markov processes. It will be an exciting challenge for practicing statisticians to apply this newfound stochastic perspective when modeling dynamic data.

How can the mathematical relations in Equation 32 be used? We give a speculative example: Consider a stochastic biophysical process X_t in stationarity and assume its stationary density $f^{\mathrm{NESS}}(\mathbf{x})$ is known. Now one carries out a measurement at time t_0 and observes $\mathbf{X}_{t_0} = \mathbf{x}_0 \pm \epsilon$. Conditioning on this information, the process is no longer stationary. Indeed, the system possesses an amount of chemical energy, which can be utilized for $t > t_0$. According to thermodynamic theory, the amount of energy is $\Psi[f(\mathbf{x}, t_0)] = -\ln\left(f^{\mathrm{NESS}}(\mathbf{x}_0)/(2\epsilon)\right)$. This result is consistent with information theory. How to calibrate this mathematical result against energy in joules and calories, however, is a challenge.

8. SUMMARY AND OUTLOOK

Biological dynamics are complex. Uncertainty is one of the hallmarks of complex behavior, either in the cause(s) of an occurred event or in the prediction of its future—modeling and predicting

weather, for example. This intuitive sense can be mathematically justified: Voigt (1981) showed that the generalized free energy Ψ defined in Equation 31 is monotonically decreasing if a dynamics is stochastic with uncertainty in the future or is deterministic but noninvertable with uncertainty in the past (i.e., many-to-one in discrete time). Ψ is conserved in one-to-one dynamics such as has been determined by differential equations! In contrast to the deterministic view of classical physics with certainty, quantitative descriptions of biological systems and processes require a statistical perspective (Qian 2013b), as demonstrated by many successful theories and discoveries from population genetics, genomics, and bioinformatics. Within single-molecule biophysics, where individual molecules are followed one at a time to study their behavior and interactions, this stochastic view is fundamental: The random motion of and interaction between molecules in time and space are necessarily described by stochastic processes. As discussed in this review, the basic laws and our understanding of statistical mechanics has naturally led to many stochastic processes that govern the behavior of the underlying single-molecule system. More importantly, our understanding and advances in stochastic-process theory have motivated new physical and chemical concepts—for example, entropy production in NESS was developed from studies of irreversible Markov processes. Statistical inference of single-molecule experimental data, including exploratory data analysis, tests of stochastic models, and estimation of model parameters, has the distinctive feature that the data are typically not the familiar i.i.d. (or independence) type. Often, the underlying stochastic-process model does not offer closed-form likelihood; even numerical evaluations are difficult in many models. Missing data, in the form of missing components/states or state aggregation, are prevalent owing to experimental limitations. There are many outstanding problems associated with stochastic model building, theoretical investigation of stochastic processes, testing of stochastic models, and estimation of model parameters. Developments in stochastic-process theory and statistical analysis of stochastic-process data will provide new modeling and data-analysis tools for biologists, chemists, and physicists. We believe these problems present great opportunities for statisticians and probabilists, not only to provide correlations and distributions, but also to determine mechanistic causality through statistical analysis.

Stochastic process is a more natural language than are classical differential equations for understanding chemical and biochemical dynamics at the level of single molecules in aqueous solutions and individual cells. It is still not widely appreciated that many of the key notions in chemistry echo important concepts in the theory of probability: transition state as the origin of a rare event, chemical potential as a form of stationary probability, a Gaussian chain as a consequence of the central limit theorem, and the potential of mean force as a manifestation of conditional probability, to name a few. All these chemical concepts have fundamental roots in statistics, though most were developed independently by chemists without explicit use of modern theory of probability and stochastic processes.

Before closing, we would like to discuss a philosophical point inevitably encountered in statistical modeling of complex dynamic data. A fundamental reason to study dynamics within classical sciences is to establish causal relations between events in the sense that modern scientific understanding demands a mechanism beyond mere statistical correlations. However, nondeterministic dynamics with random elements raises a very different kind of understanding: A force that exists on a population level may not exist on an individual level; the former is an emergent phenomenon.

Take Fick's law as an example. For a large collection of i.i.d. Brownian particles with diffusion coefficient D, their density flux clearly follows $\mathbf{J}(\mathbf{x}, t) = -D\nabla c(\mathbf{x}, t)$, where $c(\mathbf{x}, t)$ is the concentration of the particle. A net movement of the particle population is due to more particles moving from a high-concentration region to a low-concentration region than the reverse, even though every particle moves in a completely random direction. A Fickean force pushes the particle population, but this force is not acting on any one individual in the population. Therefore, this Fickean

force is a simple example of the concept of entropic force discussed in Section 6.2. In fact, because $D = k_B T/\zeta$, $\mathbf{J}(\mathbf{x}, t)$ can be expressed as $(1/\zeta)\nabla S(\mathbf{x}, t) \times c(\mathbf{x}, t)$, where $S(\mathbf{x}, t) = -k_B T \ln c(\mathbf{x}, t)$ is a form of energy if one applies Boltzmann's law in reverse.

This simple example illustrates how statistical understanding of stochastic dynamics requires an appreciation of a fundamentally novel type of law of force that has no mechanical counterpart, i.e., the notion of entropy first developed by physicists in thermodynamics. But its significance goes far beyond molecular physics, as does the second law of thermodynamics that accompanies it. In fact, we believe these concepts are firmly grounded in the domain of probability and statistics. More and deeper investigations are clearly needed.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Professor Sunney Xie for fruitful collaborations and many inspiring discussions and for sharing experimental data. Research by S.C.K. was supported in part by NIH/NIGMS grant R01GM090202.

LITERATURE CITED

Agmon N, Hopfield JJ. 1983. Transient kinetics of chemical reactions with bounded diffusion perpendicular to the reaction coordinate: intramolecular processes with slow conformational changes. *J. Chem. Phys.* 78:6947–59

Bagchi B, Blainey P, Xie XS. 2008. Diffusion constant of a nonspecifically bound protein undergoing curvilinear motion along DNA. *J. Phys. Chem. B* 112:6282–84

Ball FG, Rice JA. 1992. Stochastic models for ion channels: introduction and bibliography. *Math. Biosci.* 112:189–206

Berglund AJ. 2010. Statistics of camera-based single-particle tracking. *Phys. Rev. E* 82:011917

Bharucha-Reid AT. 1960. *Elements of the Theory of Markov Processes and Their Applications*. New York: McGraw-Hill

Bingham NH, Dunham B. 1997. Estimating diffusion coefficients from count data: Einstein-Smoluchowski theory revisited. *Ann. Inst. Stat. Math.* 49:667–79

Blainey PC, Luo G, Kou SC, Mangel WF, Verdine GL, et al. 2009. Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* 16:1224–29

Bouchaud J, Georges A. 1990. Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Phys. Rep.* 195:127–293

Braun JV, Braun RK, Muller HG. 2000. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* 87:301–14

Brenner SL, Nossal RJ, Weiss GH. 1978. Number fluctuation analysis of random locomotion: statistics of a Smoluchowski process. *J. Stat. Phys.* 18:1–18

Cai T, Munk A, Schmidt-Hieber J. 2010. Sharp minimax estimation of the variance of Brownian motion corrupted with Gaussian noise. *Stat. Sin.* 20:1011–24

Chandler D. 1987. *Introduction to Modern Statistical Mechanics*. Oxford, UK: Oxford Univ. Press

Chandrasekhar S. 1943. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.* 15:1–89

Chowdhury D. 2013. Stochastic mechano-chemical kinetics of molecular motors: a multidisciplinary enterprise from a physicist's perspective. *Phys. Rep.* 529:1–197

Delbrück M. 1940. Statistical fluctuations in autocatalytic reactions. *J. Chem. Phys.* 8:120–24

Dertinger T, Pacheco V, von der Hocht I, Hartmann R, Gregor I, Enderlein J. 2007. Two-focus fluorescence correlation spectroscopy: a new tool for accurate and absolute diffusion measurements. *ChemPhysChem* 8:433–43

Doi M, Edwards SF. 1988. *The Theory of Polymer Dynamics*. Oxford, UK: Oxford Univ. Press

Doob JL. 1942. Topics in the theory of Markoff chain. *Trans. Am. Math. Soc.* 52:37–64

Du C, Kao CL, Kou SC. 2013. Stepwise signal extraction via marginal likelihood. Preprint. **http://www.people.fas.harvard.edu/~skou/publication.htm**

Du C, Kou SC. 2012. Correlation analysis of enzymatic reaction of a single protein molecule. *Ann. Appl. Stat.* 6:950–76

Embrechts P, Maejima M. 2002. *Self-Similar Processes*. Princeton, NJ: Princeton Univ. Press

English BP, Min W, van Oijen AM, Lee KT, Luo G, et al. 2006. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nat. Chem. Bio.* 2:87–94

Esposito M, van den Broeck C. 2010. Three detailed fluctuation theorems. *Phys. Rev. Lett.* 104:090601

Feinberg M. 1989. Necessary and sufficient conditions for detailed balancing in mass action systems of arbitrary complexity. *Chem. Eng. Sci.* 44:1819–27

Finzi L, Gelles J. 1995. Measurement of lactose repressor-mediated loop formation and breakdown in single DNA molecules. *Science* 267:378–80

Fisher ME, Kolomeisky AB. 1999. The force exerted by a molecular motor. *Proc. Natl. Acad. Sci. USA* 96:6597–602

Flory PJ. 1969. *Statistical Mechanics of Chain Molecules*. New York: Wiley Intersci.

Fowler RH, Milne EA. 1925. A note on the principle of detailed balancing. *Proc. Natl. Acad. Sci. USA* 11:400–2

Fredkin DR, Rice JA. 1986. On aggregated Markov processes. *J. Appl. Prob.* 23:208–14

Ge H, Qian H. 2010. The physical origins of entropy production, free energy dissipation and their mathematical representations. *Phys. Rev. E* 81:051133

Ge H, Qian M, Qian H. 2012. Stochastic theory of nonequilibrium steady states (part II): applications in chemical biophysics. *Phys. Rep.* 510:87–118

Geva E, Skinner JL. 1998. Two-state dynamics of single biomolecules in solution. *Chem. Phys. Lett.* 288:225–29

Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–61

Gloter A, Jacod J. 2001a. Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM Prob. Stat.* 5:225–42

Gloter A, Jacod J. 2001b. Diffusions with measurement errors. II. Optimal estimators. *ESAIM Prob. Stat.* 5:243–60

Halford SE, Marko JF. 2004. How do site-specific DNA-binding proteins find their targets? *Nucleic Acid. Res.* 32:3040–52

Hotz T, Schütte O, Sieling H, Polupanow T, Diederichsen U, et al. 2012. *Idealizing ion channel recordings by jump segmentation and statistical multiresolution analysis*. Work. Pap. Appl. Math. Stat. Res. Group, Inst. Math. Stoch., Univ. Goettingen; **http://www.stochastik.math.uni-goettingen.de/preprints/IonMRC.pdf**

Jiang D-Q, Qian M, Qian M-P. 2004. *Mathematical Theory of Nonequilibrium Steady States*. Lect. Notes Math., Vol. 1833. New York: Springer

Kac M. 1959. *Probability and Related Topics in Physical Sciences*. Lect. Appl. Math., Vol. 1. New York: Interscience

Kac M. 1985. *Enigmas of Chance: An Autobiography*. New York: Harper & Row

Kelly FP. 1979. *Reversibility and Stochastic Networks*. Chichester, UK: Wiley

Kim WH. 2011. *On the behavior of the entropy production rate of a diffusion process in nonequilibrium steady state*. PhD Thesis, Univ. Washington, Seattle

Kirkwood JG. 1935. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* 3:300–13

Klafter J, Shlesinger M, Zumofen G. 1996. Beyond Brownian motion. *Phys. Today* 49:33–39

Kolomeisky AB, Fisher ME. 2007. Molecular motors: a theorist's perspective. *Annu. Rev. Phys. Chem.* 58:675–95

Kou SC. 2008a. Stochastic modeling in nanoscale biophysics: subdiffusion within proteins. *Ann. Appl. Stat.* 2:501–35

Kou SC. 2008b. Stochastic networks in nanoscale biophysics: modeling enzymatic reaction of a single protein. *J. Am. Stat. Assoc.* 103:961–75

Kou SC, Cherayil BJ, Min W, English BP, Xie XS. 2005. Single-molecule Michaelis-Menten equations. *J. Phys. Chem. B* 109:19068–81

Kou SC, Xie XS. 2004. Generalized Langevin equation with fractional Gaussian noise: subdiffusion within a single protein molecule. *Phys. Rev. Lett.* 93:180603

Kou SC, Xie XS, Liu JS. 2005. Bayesian analysis of single-molecule experimental data (with discussion). *J. R. Stat. Soc. C* 54:469–506

Kramers HA. 1940. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7:284–304

Lewis GN. 1925. A new principle of equilibrium. *Proc. Natl. Acad. Sci. USA* 11:179–83

Li G-P, Qian H. 2002. Kinetic timing: a novel mechanism for improving the accuracy of GTPase timers in endosome fusion and other biological processes. *Traffic* 3:249–55

Lu HP, Xun L, Xie XS. 1998. Single molecule enzymatic dynamics. *Science* 282:1877–82

McDunnough P. 1978. Some aspects of the Smoluchowski process. *J. Appl. Prob.* 15:663–74

McQuarrie DA. 1967. Stochastic approach to chemical kinetics. *J. Appl. Prob.* 4:413–78

Metzler R, Klafter J. 2000. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Phys. Rep.* 339:1–77

Michalet X. 2010. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E* 82:041914

Michalet X, Berglund AJ. 2012. Optimal diffusion coefficient estimation in single-particle traking. *Phys. Rev. E* 85:061916

Min W, English B, Luo G, Cherayil B, Kou SC, Xie XS. 2005a. Fluctuating enzymes: lessons from single-molecule studies. *Acc. Chem. Res.* 38:923–31

Min W, Gopich IV, English BP, Kou SC, Xie XS, Szabo A. 2006. When does the Michaelis-Menten equation hold for fluctuating enzymes? *J. Phys. Chem. B* 110:20093–97

Min W, Luo G, Cherayil B, Kou SC, Xie XS. 2005b. Observation of a power law memory kernel for fluctuations within a single protein molecule. *Phys. Rev. Lett.* 94:198302

Nagy I, Tóth J. 2012. Microscopic reversibility or detailed balance in ion channel models. *J. Math. Chem.* 50:1179–99

Perrin J-B. 1916. *Atoms*, transl. DL Hammick, D. van Nostrand. New York: Kessinger

Qian H. 1990. On the statistics of fluorescence correlation spectroscopy. *Biophys. Chem.* 38:49–57

Qian H. 1997. A simple theory of motor protein kinetics and energetics. *Biophys. Chem.* 67:263–67

Qian H. 2000. A mathematical analysis of the Brownian dynamics of DNA tether. *J. Math. Biol.* 41:331–40

Qian H. 2001. Mathematical formalism for isothermal linear irreversibility. *Proc. R. Soc. A.* 457:1645–55

Qian H. 2002. Equations for stochastic macromolecular mechanics of single proteins: equilibrium fluctuations, transient kinetics and nonequilibrium steady state. *J. Phys. Chem. B* 106:2065–73

Qian H. 2003. Fractional Brownian motion and fractional Gaussian noise. In *Processes with Long-Range Correlations: Theory and Applications*, Vol. 621, ed. G Rangarajan, MZ Ding, pp. 22–33. New York: Springer

Qian H. 2005. Cycle kinetics, steady-state thermodynamics and motors: a paradigm for living matter physics. *J. Phys. Condens. Matt.* 17:S3783–94

Qian H. 2008. Cooperativity and specificity in enzyme kinetics: a single-molecule time-based perspective (mini review). *Biophys. J.* 95:10–17

Qian H. 2013a. A decomposition of irreversible diffusion processes without detailed balance. *J. Math. Phys.* 54:053302

Qian H. 2013b. Stochastic physics, complex systems and biology. *Quant. Biol.* 1:50–53

Qian H, Elson EL. 1999. Quantitative study of polymer conformation and dynamics by single-particle tracking. *Biophys. J.* 76:1598–605

Qian H, Elson EL. 2002. Single-molecule enzymology: stochastic Michaelis-Menten kinetics. *Biophys. Chem.* 101:565–76

Qian H, Elson EL. 2004. Fluorescence correlation spectroscopy with high-order and dual-color correlation to probe nonequilibrium steady-states. *Proc. Natl. Acad. Sci. USA* 101:2828–33

Qian H, Qian M. 2000. Pumped biochemical reactions, nonequilibrium circulation, and stochastic resonance. *Phys. Rev. Lett.* 84:2271–74

Qian H, Qian M, Tang X. 2002. Thermodynamics of the general diffusion process: time reversibility and entropy production. *J. Stat. Phys.* 107:1129–41
Qian H, Raymond GM, Bassingthwaighte JB. 1999. Stochastic fractal behaviour in concentration fluctuation and fluorescence correlation spectroscopy. *Biophys. Chem.* 80:1–5
Qian H, Sheetz MP, Elson EL. 1991. Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. *Biophys. J.* 60:910–21
Qian M-P, Qian M, Gong G-L. 1991. The reversibility and entropy production of Markov processes. *Contemp. Math.* 118:255–61
Reuter GEH. 1957. Denumerable Markov processes and the associated contraction semigroups on ℓ. *Acta Math.* 97:1–46
Ridgeway WK, Millar DP, Williamson JR. 2012. The spectroscopic basis of fluorescence triple correlation spectroscopy. *J. Phys. Chem.* 116:1908–19
Rigler R, Elson EL. 2001. *Fluorescence Correlation Spectroscopy: Theory and Applications*. Chem. Phys. Ser., Vol. 65. New York: Springer
Rothberg BS, Magleby KL. 2001. Testing for detailed balance (microscopic reversibility) in ion channel gating. *Biophys. J.* 80:3025–26
Ruben H. 1963. The estimation of a fundamental interaction parameter in an emigration-immigration process. *Ann. Math. Stat.* 34:238–59
Ruben H. 1964. Generalized concentration fluctuations under diffusion equilibrium. *J. Appl. Prob.* 1:47–68
Sakmann B, Neher E, eds. 2009. *Single-Channel Recording*. New York: Springer. 2nd ed.
Saxton MJ, Jacobson K. 1997. Single-particle tracking: applications to membrane dynamics. *Annu. Rev. Biophys. Biomol. Struct.* 26:373–99
Schafer DA, Gelles J, Sheetz MP, Landick R. 1991. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature* 352:444–48
Schellman JA. 1980. The flexibility of DNA: I. Thermal fluctuations. *Biophys. Chem.* 11:321–28
Schenter GK, Lu HP, Xie XS. 1999. Statistical analysis and theoretical models of single-molecule enzymatic dynamics. *J. Phys. Chem. A* 103:10477–88
Schurr JM. 1979. The one-dimensional diffusion coefficient of proteins absorbed on DNA. Hydrodynamic considerations. *Biophys. Chem.* 9:413–14
Schuss Z. 2010. *Theory and Applications of Stochastic Processes: An Analytical Approach*. New York: Springer
Schwille P, Meyer-Almes FJ, Rigler R. 1997. Fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution. *Biophys. J.* 72:1878–86
Seifert U. 2012. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.* 75:126001
Sisan DR, Yarar D, Waterman CM, Urbach JS. 2010. Event ordering in live-cell imaging determined from temporal cross-correlation asymmetry. *Biophys. J.* 98:2432–41
Slutsky M, Mirny LA. 2004. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys. J.* 87:4021–35
Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. 2005. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. B* 67:91–108
Tu Y. 2008. The nonequilibrium mechanism for ultrasensitivity in a biological switch: sensing by Maxwell's demons. *Proc. Natl. Acad. Sci. USA* 105:11737–41
Voigt J. 1981. Stochastic operators, information, and entropy. *Commun. Math. Phys.* 81:31–38
Wang H, Qian H. 2007. On detailed balance and reversibility of semi-Markov processes and single-molecule enzyme kinetics. *J. Math. Phys.* 48:013303
Wang J, Wolynes PG. 1995. Intermittency of single molecule reaction dynamics in fluctuating environments. *Phys. Rev. Lett.* 74:4317–20
Wax N, ed. 1954. *Selected Papers on Noise and Stochastic Processes*. New York: Dover
Weber SC, Thompson MA, Moerner WE, Spakowitz AJ, Theriot JA. 2012. Analytical tools to distinguish the effects of localization error, confinement, and medium elasticity on the velocity autocorrelation function. *Biophys. J.* 102:2443–50
Wiener N. 1966. *Nonlinear Problems In Random Theory*. Boston, MA: MIT Press

Witkoskie JB, Cao J-S. 2006. Testing for renewal and detailed balance violations in single-molecule blinking processes. *J. Phys. Chem. B* 110:19009–17
Xie XS. 2001. Single molecule approach to enzymology. *Single Mol.* 4:229–36
Xie XS, Lu HP. 1999. Single-molecule enzymology. *J. Biol. Chem.* 274:15967–70
Yang H, Luo G, Karnchanaphanurach P, Louise T-M, Rech I, et al. 2003. Protein conformational dynamics probed by single-molecule electron transfer. *Science* 302:262–66
Yao YC. 1988. Estimating the number of change-points via Schwarz' criterion. *Stat. Prob. Lett.* 6:181–89
Zhang X-J, Qian H, Qian M. 2012. Stochastic theory of nonequilibrium steady states and its applications (part I). *Phys. Rep.* 510:1–86
Zwanzig R. 2001. *Nonequilibrium Statistical Mechanics*. New York: Oxford Univ. Press

Annual Review of Statistics and Its Application

Volume 1, 2014

Contents

What Is Statistics?
Stephen E. Fienberg 1

A Systematic Statistical Approach to Evaluating Evidence from Observational Studies
David Madigan, Paul E. Stang, Jesse A. Berlin, Martijn Schuemie, J. Marc Overhage, Marc A. Suchard, Bill Dumouchel, Abraham G. Hartzema, and Patrick B. Ryan 11

The Role of Statistics in the Discovery of a Higgs Boson
David A. van Dyk 41

Brain Imaging Analysis
F. DuBois Bowman 61

Statistics and Climate
Peter Guttorp 87

Climate Simulators and Climate Projections
Jonathan Rougier and Michael Goldstein 103

Probabilistic Forecasting
Tilmann Gneiting and Matthias Katzfuss 125

Bayesian Computational Tools
Christian P. Robert 153

Bayesian Computation Via Markov Chain Monte Carlo
Radu V. Craiu and Jeffrey S. Rosenthal 179

Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models
David M. Blei 203

Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues
Martin J. Wainwright 233

High-Dimensional Statistics with a View Toward Applications in Biology
Peter Bühlmann, Markus Kalisch, and Lukas Meier 255

Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data
Kenneth Lange, Jeanette C. Papp, Janet S. Sinsheimer, and Eric M. Sobel 279

Breaking Bad: Two Decades of Life-Course Data Analysis in Criminology, Developmental Psychology, and Beyond
Elena A. Erosheva, Ross L. Matsueda, and Donatello Telesca 301

Event History Analysis
Niels Keiding 333

Statistical Evaluation of Forensic DNA Profile Evidence
Christopher D. Steele and David J. Balding 361

Using League Table Rankings in Public Policy Formation: Statistical Issues
Harvey Goldstein 385

Statistical Ecology
Ruth King 401

Estimating the Number of Species in Microbial Diversity Studies
John Bunge, Amy Willis, and Fiona Walsh 427

Dynamic Treatment Regimes
Bibhas Chakraborty and Susan A. Murphy 447

Statistics and Related Topics in Single-Molecule Biophysics
Hong Qian and S.C. Kou 465

Statistics and Quantitative Risk Management for Banking and Insurance
Paul Embrechts and Marius Hofert 493