Statistical Science
2020, Vol. 35, No. 1, 75–91
https://doi.org/10.1214/19-STS752

Statistical Methodology in Single-Molecule Experiments

Chao Du and S. C. Kou

Abstract. Toward the last quarter of the 20th century, the emergence of single-molecule experiments enabled scientists to track and study individual molecules' dynamic properties in real time. Unlike macroscopic systems' dynamics, those of single molecules can only be properly described by stochastic models even in the absence of external noise. Consequently, statistical methods have played a key role in extracting hidden information about molecular dynamics from data obtained through single-molecule experiments. In this article, we survey the major statistical methodologies used to analyze single-molecule experimental data. Our discussion is organized according to the types of stochastic models used to describe single-molecule systems as well as major experimental data collection techniques. We also highlight challenges and future directions in the application of statistical methodologies to single-molecule experiments.

Key words and phrases: Autocorrelation, continuous-time Markov chain, diffusion process, heterogeneity, hidden Markov model, molecular dynamics.

1. THE ARRIVAL OF SINGLE-MOLECULE EXPERIMENTS AND THE ROLE OF STATISTICAL METHODOLOGY

The modern concept of the molecules was introduced in the 19th century. Since then, questions about how interactions among molecules dictate macroscopic systems' properties have become areas of key scientific inquiry. Although the foundations of many scientific theories, such as statistical mechanics, were firmly grounded at the molecular level, knowledge of molecular dynamics was mostly derived from data obtained in ensemble-level experiments (i.e., experiments involving many molecules). Such data provided us information on the average properties of molecules rather than individual molecules' dynamics. Starting in the late 20th century, scientists have developed new techniques (see the Supplementary Material, Du and Kou, 2020) to manipulate and visualize single molecules in liquid solutions. These advances heralded the era of single-molecule experiments and significantly advanced our understanding of microscopic systems.

To analyze single-molecule experimental data, advanced statistical methodologies are necessary. First, unlike traditional technological advancements that often improve the signal-to-noise ratio, single-molecule experiments reveal an additional layer of uncertainty hidden in conventional ensemble-level experiments. The microscopic realm is intrinsically stochastic and can only be sufficiently described with probability models. Many single-molecule experiments also rely on recording fluorescence emission, a stochastic process governed by quantum mechanics. Thus, even if we repeat a single-molecule experiment perfectly under the exact same conditions, we cannot reproduce the same data. In this regard, a key feature of single-molecule data analysis is to learn from noisy stochastic data. Sophisticated statistical approaches are therefore required to analyze the distribution of observed data and infer hidden information.

Second, single-molecule experimental data analysis must account for a considerable degree of heterogeneity over time and across molecules. A single molecule is subject to constant fluctuation that can affect its properties. A stochastic model with fixed parameters may be insufficient to model such phenomena. It is often necessary to allow the parameters or even the model to fluctuate over time. Many single-molecule experiments also record data originating from multiple molecules that exhibit distinct characteristics beyond the explanatory power of a simple stochastic model. Suitable statistical methods are

Chao Du is Assistant Professor in Statistics, University of Virginia, Halsey Hall, Charlottesville, Virginia 22903, USA (e-mail: cd2wb@virginia.edu). S. C. Kou is Professor of Statistics and Professor of Biostatistics, Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: kou@stat.harvard.edu).

therefore necessary to infer molecules' common traits and individual characteristics.

Third, despite the technological advancements, modern single-cell experiments often only illuminate a small part (e.g., the surface) of the system under investigation. Many single-molecule experiments are conducted to validate or improve scientific hypotheses. However, more often than not, the observed data do not contain sufficient information to directly examine these hypotheses. Even when suitable stochastic models are proposed, the data generally only contain partial output of these models. Therefore, sophisticated statistical approaches are necessary to bridge the gap between the observed data and the questions of interest.

This clearly evidences why statistical approaches have been used to analyze experimental data since the emergence of single-molecule experiments. This union has only grown stronger in recent decades as modern statistical methodologies have gradually been adopted in this field. In this article, we review several major statistical methods that have been used to analyze single-molecule experimental data in recent decades. Our review is mainly organized according to the stochastic models used for modeling molecular systems, which are usually tied to particular experimental techniques. We will explore the following three areas: analyzing the motion of a single molecule modeled as diffusion process, studying the fluctuation of molecular number within a small volume based on the autocorrelation function of intensity signal, and inferring the dynamic mechanism of a single molecule modeled by continuous-time Markov chains.

2. MOTION OF A SINGLE MOLECULE AND DIFFUSION PROCESS

The famous Brownian motion refers to the seemingly chaotic movement of small particles in a solution. In 1905, Albert Einstein proposed a theoretical model for the Brownian motion, which links the distribution of small particles' trajectories to the properties of the particles and the surrounding environment. In the 1980s, the development of the single-particle tracking (SPT) technique allowed scientists to record the positions of single particles at the speed of a few milliseconds per frame (see a review by Saxton and Jacobson, 1997). In 1996, the obstacle of tracking a single molecule's locations was overcome using fluorescence spectroscopy (Schmidt et al., 1996). The observed positions of molecules can be used to reconstruct single molecules' trajectories over time (see the Supplementary Material, Du and Kou, 2020), which allows scientists to study the properties of biomolecules and the structure of cellular systems.

2.1 Free Diffusion and Diffusion Constant

The displacement $x(t)$ of a small particle in one-dimensional space without the interference of an external field can be modeled by the integrated Ornstein-Uhlenbeck process, which is usually approximated by a scaled standard Wiener process (see the Supplementary Material, Du and Kou, 2020). This approximation is referred to as Brownian motion or free diffusion in the biophysics literature. For free diffusion, the second moment of $x(t)$ follows the Einstein-Smoluchowski relation:

$$E[x^2(t)] = 2Dt. \tag{1}$$

A similar relationship holds in higher-dimensional space: the second moment of $\|x(t)\|$ is $4Dt$ in two dimensions and $6Dt$ in three dimensions. The constant parameter D depends on the temperature, the solution's viscosity, and the particle's radius, and is known as the diffusion constant (see the Supplementary Material, Du and Kou, 2020). The estimated diffusion constant can shed light on the particle's properties and its interaction with the surrounding environment (Blainey et al., 2009).

The mean square displacements (MSDs) of a single particle under free diffusion are natural candidates for estimating D. In typical SPT-based experiments, the positions of a molecule are measured at discrete time points with constant increments. If we use $x_1, \dots, x_n$ to denote the molecule's measured positions at time $t_1, \dots, t_n$ ($\Delta t = t_{i+1} - t_i$), the MSD of this molecule after time lag $k\Delta t$ can be estimated by averaging all displacements with time lag $k\Delta t$ (Qian, Sheetz and Elson, 1991):

$$\hat{\rho}_k = \frac{1}{n-k} \sum_{i=1}^{n-k} (x_{i+k} - x_i)^2, \quad k = 1, 2, \dots. \tag{2}$$

A simple regression of the estimated MSD $\hat{\rho}_k$ against k can be used to estimate D. Because $\hat{\rho}_k$ is subjected to larger variation for larger k, weighted regression is the preferred approach (Qian, Sheetz and Elson, 1991; Saxton, 1997). Alternatively, MSDs can also be estimated by averaging all nonoverlapping displacements, which yields slightly greater variance (Saxton, 1997).

Without measurement error, the aforementioned method of estimating the diffusion constant can reach the usual convergence rate of $n^{-1/2}$. In practice, the measurements of molecules' locations are subjected to the static error when measuring the location of an immobilized molecule, as well as the motion blur error due to the movement during Δt, and the theoretical rate of convergence would be much slower (at $n^{-1/4}$) (Gloter and Jacod, 2001). Methods of estimating diffusion constants in the presence of measurement errors include the maximum likelihood method (Berglund, 2010), the optimized least-squares fit method (Michalet, 2010) and the covariance-based estimator (Vestergaard, Blainey and Flyvbjerg, 2014).

2.2 Inference Beyond Free Diffusion

Anomalous diffusion. In many physical and biological systems, instead of being a linear function of time, the MSD of the molecular trajectory obeys the following general power law:

$$E[x^2(t)] = Dt^\alpha, \tag{3}$$

which is known as anomalous diffusion. Various theoretical frameworks have been proposed to model anomalous diffusion (for an in-depth review, see Metzler et al., 2014). Notable theoretical models include the fractional models (Metzler and Klafter, 2000) and the models based on generalized Langevin equations with fractional Gaussian noise (Kou and Xie, 2004; Kou 2008a; also see the Supplementary Material, Du and Kou, 2020).

Both of the anomalous exponent α and the (generalized) diffusion constant D can be estimated by a simple regression model with both MSD and time at log-scale. More sophisticated approaches would take measurement errors (Kepten, Bronshtein and Garini, 2013; Sikora et al., 2017a, 2017b) as well as the uncertainty in MSD estimates (Kepten et al., 2015) into consideration. The asymptotic properties of the MSD-based estimators in anomalous diffusion processes are largely unsettled. When the anomalous diffusion process is modeled as fractional Brownian motion, progress has been made (Sikora et al., 2017a, 2017b).

Directed diffusion. The motion of a molecule subjected to a constant directed drift with velocity V is known as directed diffusion. Within a cell, the presence of directed diffusion often indicates an active molecule transport mechanism. The MSD of directed diffusion is

$$E[x^2(t)] = 2Dt + V^2t^2, \tag{4}$$

which is quadratic in time (Qian, Sheetz and Elson, 1991). Moreover, molecular trajectory under directed diffusion exhibits a clear tendency toward the drift's direction. For this reason, statistics that reflect trajectories' asymmetry are often used alongside MSD analysis to identify directed diffusion and estimate the drift velocity (Saxton, 1994; Huet et al., 2006).

Confined diffusion. When a molecule's motion is restricted within a confined space imposed by cellular structure, confined diffusion occurs. Confined diffusion can be described as a diffusion process that occurs in a finite space with reflection boundaries. Although the exact formula of the MSD of confined diffusion depends on the characteristics of the confined space, the MSD curve will not increase indefinitely and will arrive and remain at a plateau after sufficiently large t. This characteristic can be used to identify confinement in diffusion processes and estimate the size of the confined space (Kusumi, Sako and Mutsuya, 1993; Jeon and Metzler, 2010; Clausen and Lagerholm, 2013).

2.3 Inferring Heterogeneity in Single-Molecule Trajectories

A homogeneous diffusion process with constant parameters is insufficient to describe the heterogeneity observed in single-molecule trajectories. The characteristics of the diffusion process usually depend on the molecule's properties, including its size, mass, and structure, which may vary between molecules and fluctuate during a molecule's lifetime. As the molecule traverses the complex landscape within a cell, environmental factors can change as well. Consequently, heterogeneity may not only appear across trajectories but may also manifest over the course of the same trajectory. From the modeling perspective, observed trajectories can switch between diffusion states characterized by distinctive parameters or even different types of diffusion processes. From the inference perspective, statistical methods are necessary to identify and categorize these underlying diffusion states and estimate the transitions between diffusion states or diffusion processes.

2.3.1 *Heterogeneity across trajectories.* When heterogeneity mainly appears as variation between trajectories, such as in the case of multiple short trajectories, each trajectory can be described using a homogeneous diffusion process drawing from a collection of diffusion states. Analyzing heterogeneity across trajectories is therefore a matter of clustering or classification.

The MSD curve, based on its distinctive shapes in various diffusion models, is often used as the basis of such analysis. The estimated MSD curve can provide direct clues to recognizing stationary, free, anomalous, confined, and directed diffusion (Kusumi, Sako and Mutsuya, 1993; Suh et al., 2007). A systematic approach to classifying multiple trajectories can be developed within a Bayesian framework that combines the subjective prior belief of diffusion models and the likelihood of MSDs (Monnier et al., 2012). In this approach, the joint distribution of MSDs at different time points is assumed to be multivariate Gaussian with a covariance matrix derived from mathematical models. Random forest can be applied to classify multiple trajectories using features derived from MSDs (Wagner et al., 2017). Trajectories can also be classified from the distributions of observed displacements directly (Koo et al., 2015).

2.3.2 *Heterogeneity over time.* When heterogeneity appears within the same trajectory, the first priority is to identify the transitions between different diffusion states and discover the transitions' patterns. Because MSDs are generally estimated by averaging over the whole trajectory, MSD-based analysis has limited use when solving such problems. Approaches that can extract information from local segments are often necessary. Here we will focus on how to analyze heterogeneity in a single observed trajectory. Nonetheless, many approaches discussed in this section can also be used to study multiple statistically independent trajectories.

Fluctuation of diffusion constants in free diffusion. If the molecule follows free Brownian diffusion through its course of movement, heterogeneity over time can only appear as fluctuation in diffusion constants. In this scenario, each diffusion state corresponds to a distinct diffusion constant. Given how the diffusion constant changes over time, the likelihood function of the observed trajectory can be easily obtained as the displacements are independently Gaussian distributed. Therefore, hidden Markov models with known numbers of states are widely used to detect changes in diffusion constants (Das, Cairo and Coombs, 2009; Chung et al., 2010; Ott, Shai and Haran, 2013; Slator, Cairo and Burroughs, 2015). If the number of diffusion states is unknown, the posterior distribution of the number of states (Persson et al., 2013) or model selection criterion such as BIC (Koo et al., 2015) can be used to determine the number of states. Other approaches also exist, such as modeling the overall distribution of all displacements as a mixture distribution in which each component corresponds to a particular diffusion constant (Bosch, Kanger and Subramaniam, 2014) or applying likelihood-ratio-based tests to detect transitions between diffusion constants as in a typical change-point problem (Yin, Song and Yang, 2018).

Detecting changes in the modes of diffusion. More generally, heterogeneity within the same trajectory can manifest when transitions occur between different diffusion processes. A common problem in this area is identifying segments of confined diffusion in a trajectory dominated by free diffusion. This problem arises from studying the membrane protein, whose interaction with the membrane may cause it to become temporally trapped in a small compartment. Analyzing this mechanism can then provide valuable information about the membrane structure and about the interactions between the membrane and the protein.

One approach for identifying transient confinement is to rely on suitable segment-level statistics to distinguish confined and free diffusion. In free diffusion, a molecule's motion is not restricted by a physical boundary, so the molecule tends to move further. Therefore, transition from free diffusion to confined diffusion can be detected by tracking the estimated probability of a molecule reaching maximum displacement within Δt (Saxton, 1993; Simson, Sheets and Jacobson, 1995). This approach can be further improved by considering the variation in the maximum displacement (Meilhac et al., 2006). A drawback of this approach is that the diffusion constant must be estimated prior to transition identification. This can be avoided by relying on the so-called packing coefficient, which is defined as the ratio of the observed segment's length to the area of convex hull containing the observed segment (Renner et al., 2017).

The aforementioned approaches depend on the choice of the length of local segment to construct testing statistics. Other approaches, such as hidden Markov models, have been used to identify transient confinement in recent years. To establish analytical likelihood, confined diffusion is often modeled as free diffusion in the presence of a potential well. Transitions between the confined and free diffusion, as well as other parameters of interest, such as diffusion constant and the strength of the potential well, can then be inferred using particle filtering (Bernstein and Fricks, 2016) or MCMC algorithm (Slator and Burroughs, 2018) under a two-state hidden Markov model framework.

Transitions in diffusion states can occur between other diffusion processes as well. Within cells, the transportations of molecules often involve the stochastic transitions between free and directed diffusion. Such transitions can be detected based on the asymmetry in molecular trajectories (Huet et al., 2006), local MSD estimated based on short time window (Arcizet et al., 2008), or a hidden Markov model with a Bayesian model selection method (Monnier et al., 2015). More generally, by modeling the joint distribution of observed displacements, transitions between free, confined, and anomalous diffusions and immobile states can also be estimated using likelihood approaches (Koo and Mochrie, 2016).

3. FLUCTUATION IN MOLECULAR NUMBERS AND AUTOCORRELATION FUNCTION

The diffusion of molecules and chemical reactions can cause the number of molecules to fluctuate within a given volume. By analyzing the fluctuation in the number of molecules over time, we can extract useful information such as diffusion constants or chemical reaction rates. Fluorescence correlation spectroscopy (FCS), which was developed almost half a century ago (Magde, Elson and Webb, 1972), has been used for this purpose. FCS utilizes a confocal microscope to record the arrivals of a single photon stream emitted by fluorescent molecules within a tiny detection volume upon the excitation of a focused laser beam. Raw photon arrival data can then be processed to represent the overall strength of fluorescence originating from all molecules at time t: the so-called fluorescence intensity signal $F(t)$. Although FCS does not actively track individual molecules, modern FCS experiments can detect the change caused by a single molecule (see the review by Elson, 2011).

3.1 Autocorrelation Function of Fluorescence Intensity

Under the equilibrium assumption, fluorescence intensity $F(t)$ is a stationary stochastic process with autocorrelation function $G(\tau)$:

$$G(\tau) = \frac{\mathrm{Cov}(F(0), F(\tau))}{[E(F(0))]^2}, \tag{5}$$

where the expectation is computed over the whole time course. For example, the first moment of intensity $E(F(0))$ can be computed as $\frac{1}{T}\int_0^T F(t)\,dt$.

The autocorrelation function of the $F(t)$ depends on the cause of the underlying system's fluctuation in molecular number. In many common scenarios, the analytic formula of $G(\tau)$ can be derived (Elson and Magde, 1974) so that the key parameters can be estimated based on the observed autocorrection. For instance, if there is only a single fluorescence molecular species within the detection volume and its copy number follows Poisson distribution, the autocorrelation function at 0 equals the inverse of the average number of molecules. If fluctuation in the molecular number is caused by the free diffusion of molecules that move in or out of the detection volume, then $G(\tau)/G(0) = (1 + \tau/\tau_D)^{-1}$ with $\tau_D \propto D^{-1}$. If the fluctuation in molecular number is caused by a chemical reaction, the autocorrelation function will explicitly depend on the reaction rate. If more than one fluorescent molecular species exist in the system, the recorded intensity $F(t)$ equals the sum of fluorescence intensities from each species, and the autocorrelation function will provide a clue to the presence of multiple molecular species.

3.2 Inference with Autocorrelation Function

A common way to analyze FCS data is to fit the theoretical autocorrelation function to the observed autocorrelation curve. To properly account for noise in the autocorrelation curve, χ^2 statistics, the sum of the squared errors between the observed and fitted values, normalized by the variances, can be used as the object function (Koppel, 1974; Meseth et al., 1999). The variance of correlation can be computed using approximated equations (Koppel, 1974) or estimated from multiple signals (Wohland, Rigler and Vogel, 2001).

This approach is useful for choosing competing models. Although a direct comparison of the fitted and observed autocorrelation curves can be sufficient to rule out underfitted models (Brock, Hink and Jovin, 1998; Gennerich and Schild, 2000), choosing among models that all fit the data well requires a more sophisticated approach. Reduced χ^2 statistics that penalizes model complexity have been used to construct an F-test for model comparision (Meseth et al., 1999). By approximating the likelihood of autocorrelation functions with multivariate Gaussian distribution, marginal likelihoods or posterior probabilities can also be used for such a purpose (He, Guo and Bathe, 2012; Sun et al., 2015).

The observed fluorescence intensity signal depends on the concentration and the brightness of molecules in the detection volume. When multiple molecular species with various concentrations and brightness levels are present, a conventional approach based on the autocorrelation function alone can be insufficient for distinguishing molecular species. In this scenario, it is necessary to consider higher-order moments of intensity signal (Qian and Elson, 1990a, 1990b) or the higher-order correlation functions (Melnykov and Hall, 2009; Wu et al., 2016).

3.3 Photon Count Histogram

The autocorrelation function only utilizes the fluorescence intensity signal's first two normalized moments. To extract more information, the full distribution of the photon count—the number of photons detected over a given time interval—is also used in analysis. For instance, the stationary probability of photon count n can be expressed as $P(n) = \sum_{m=0}^{\infty} G(n|m)H(m)$, where $H(m)$ is the stationary probability of the number of molecules m, and $G(n|m)$ represents the conditional probability of the number of photons generated from m molecules. Under suitable assumptions, the formula of $P(n)$ can be derived and used to fit the observed histogram of the photon count. This approach is known as photon count histogram (PCH).

A primary advantage of the PCH approach over the autocorrelation function is that it accounts for both concentration and brightness of molecular species. Specifically, $H(n)$ depends on the concentration of the molecules, and $G(n|m)$ depends on the brightness of the molecules. Due to the normalization, the autocorrelation function does not contain much information on molecular brightness. This fact makes PCH a particularly valuable tool for distinguishing multiple molecular species with distinct concentrations and brightness levels (Chen et al., 1999; Müller, Chen and Gratton, 2000). PCH can be generalized to two-dimensional space when two detectors are used (Kask et al., 2000). The cumulants of photon count can also serve as alternatives to PCH (Müller, 2004; Wu and Müller, 2005).

In practice, the observed PCH is recreated from the photon counts recorded at successive time intervals, which are effectively treated as i.i.d. observations. Although photon counts measured at different times follow the same marginal distribution when the stationary assumption holds, they are correlated. Therefore, if the primary goal is to capture the temporal dynamics of the underlying system, the autocorrelation function is still the preferred method.

4. SINGLE-MOLECULE DYNAMICS AND CONTINUOUS-TIME MARKOV CHAIN

Another major focus of single-molecule experiments is the study of chemical kinetics and the conformational dynamics of single molecules. At the molecular level, chemical reactions are triggered by random collisions between molecules. To complete a chemical reaction, the molecules involved must often go through intermediate steps that are generally hidden at the macroscopic level.

Moreover, the functions and properties of a molecule depend on its conformational structure, which is subject to fluctuation and may undergo dramatic changes under suitable conditions. Such conformational dynamics can be essential for the proper functions of molecules and can be observed and studied in single-molecule experiments. To study the chemical kinetics and the conformational dynamics of single molecules, we need to monitor the activity of a single molecule in real time. In recent years, many experimental techniques have been developed for tracking the dynamics of a single molecule through indirect means. Notable methods include patch clamp recording and fluorescence spectroscopy.

Patch clamp recording (Neher and Sakmann, 1976; see also the review by Sakmann, 2013) was developed to study the mechanism of ion channels, protein molecules on cell membranes that serve as gates for the ion current across the membrane. Through conformational changes in the protein structure, an ion channel can switch between open and closed states to enable or block the ion current. The patch clamp recording method can record the strength of the ion current passing through a small area of membrane (which may contain a single or a few ion channels) over time. The changes in the levels of observed signals can be used to determine the dynamics of the conformational state of the ion channel(s) in the recording area.

In fluorescence spectroscopy, time-varying signals can be obtained by recording the intensity of a photon stream that originates from one or several fluorescence tags attached to a single molecule (see the review by Orrit, Ha and Sandoghdar, 2014). With the help of a single-photon-counting detector, experimental devices can record each photon's arrival time to generate the so-called time-stamped photon sequence. However, analysis is usually conducted using the condensed time-binned photon intensity signal, in which each observation represents the total number of photons collected by the detector during a short time window. One advantage of the time-binned intensity signal compared to the time-stamped data is that it does not require a highly precise single-photon-counting detector and can potentially reduce the noise introduced by the intrinsic stochasticity of photon emission processes through aggregation.

Time-stamped data and time-binned signals reflect how fluorescence intensity changes over time, which is usually tied to the underlying molecules' reaction or conformational state. Some molecules have naturally fluorescent sites that can be turned on or off based on certain conformational changes. By tracking how the fluorescence signal varies, we can infer the molecule's conformational dynamic (Lu, Xun and Xie, 1998). The FRET microscopy (Ha et al., 1996) simultaneously records the fluorescence signals emitted by two fluorophores, a donor and an acceptor, attached to the same molecule. The acceptor's fluorescence intensity depends on the distance between the fluorophores. Thus, the FRET ratio, which is the ratio of the acceptor's intensity to the total intensity of the acceptor and the donor, allows us to measure the relative physical distance between two parts of the host molecule (tagged by the donor and acceptor) and collect information about its conformational dynamics.

Signals from patch clamp recording and fluorescence spectroscopy, especially the time-binned data, are time series that share a common trait: Their signal may remain at a relatively stable level (with noise) for a short time but switch stochastically between different levels in the long run. If the variation in the stable level is removed, we can obtain an idealized version of such a signal: a step function over time that can take a few distinct values.

4.1 Modeling Molecular Dynamics

A common way to model the dynamics of single molecules is to use a continuous-time Markov chain (CTMC) with finite state space. Under this model, each state refers to a stable conformational state of the molecule or a particular stage of the chemical reaction involving the molecule, and the molecule can stochastically switch between various states.

Let us assume that the dynamic of a molecule can be modeled with a CTMC with K different states $\mathbf{S} = \{S_1, S_2, \dots, S_K\}$ and the generator matrix $Q = \{q_{ij}\}$. The dwell time of the molecule in state S_i would follow an exponential distribution with rate $-q_{ii} = \sum_{j \neq i} q_{ij}$. Whenever the transition out of the current state S_i occurs, the molecule will switch into one of the other states S_j $(j \neq i)$ with a probability proportional to q_{ij}.

Example: *Enzymatic reaction*. In the famous Michaelis–Menten model of a catalytic reaction, an enzyme molecule E binds with the substrate molecule S to form a complex ES. The complex can either dissociate back into the enzyme and substrate molecule or undergo the catalytic process to transform the substrate into product P and release the enzyme molecule. The released enzyme E^0 would then return to the initial state E for another catalytic circle. This model can be summarized as follows (English et al., 2006):

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\rightarrow} P + E^0, \quad E^0 \overset{\delta}{\rightarrow} E. \tag{6}$$

Here the enzyme molecule can switch among three states: free enzyme E, complex ES, and the released state E^0. With a single enzyme molecule, the transitional rates among different states depend only on the chemical kinetic rate parameters k_1, k_{-1}, k_2, δ and the concentration of substrates, $[S]$. The generator matrix Q can be written as

$$\begin{array}{c|ccc} & E & ES & E^0 \\ \hline E & -k_1[S] & k_1[S] & 0 \\ ES & k_{-1} & -k_{-1} - k_2 & k_2 \\ E^0 & \delta & 0 & -\delta \end{array}.$$

Example: *Ion channel.* The del Castillo–Katz model (Del Castillo and Katz, 1957) models the opening and closing of the ion channel with a two-step process. Let us use T to represent the closed state of the ion channel. To transit into the open state, T needs to bind with agonist A first to form an intermediate compound AT. At state AT, the channel remains closed until the transition into the open state AR occurs. Similarly, the opened ion channel can be closed through the unbinding of agonist A in a reversed two-step process. In summary, the ion channel can switch between two closed states T and AT and one open state AR. This model, along with its generator matrix, is listed below (Colquhoun and Hawkes, 1981):

$$T \underset{k_{-1}}{\overset{k_{+1}}{\rightleftarrows}} AT \underset{\beta}{\overset{\alpha}{\rightleftarrows}} AR,$$

$$(7)\qquad \begin{array}{c} \\ T \\ AT \\ AR \end{array} \begin{array}{c} \begin{array}{ccc} T & AT & AR \end{array} \\ \left(\begin{array}{ccc} -k_{+1}[A] & k_{+1}[A] & 0 \\ k_{-1} & -k_{-1}-\alpha & \alpha \\ 0 & \beta & -\beta \end{array} \right) \end{array}.$$

A realization of the trajectory of a CTMC over $[0, T]$ can be denoted as $x(t)$ where $x(t) = k$ if the system is in the state S_k at time t. Given the full trajectory of $x(t)$ over $[0, T]$, its likelihood is the product of the exponential densities of sojourn times in successive states and the multinomial probabilities of the transitions between states. In practice, instead of observing the full trajectory of $x(t)$, we can take only indirect measurements $\mathbf{y} = (y(t_1), y(t_2), \ldots, y(t_n))$ at discrete times $0 = t_1 < t_2 < \cdots < t_n = T$. The distributions of $y(t_i)$ are usually modeled with distributional family F where the parameter values at time t_i are determined by $x(t_i)$:

$$y(t_i) \sim F(\theta_{x(t_i)}), \quad E\bigl(y(t_i)\bigr) = \mu_{x(t_i)}.$$

It is also customary to assume that as long as the parameter values are different the means of the measurements will differ as well. Thus, we can expect that the observed signal, although it can be noisy, would roughly follow a stepwise pattern.

In principle, the estimated mean signal $\hat{\boldsymbol{\mu}}$ would allow us to determine the states of the single molecule at the times of measurement and estimate the change-points (times when the transitions between states occur). Such information would allow us to reconstruct $x(t)$ over $[0, T]$, infer the parameters in the generator matrix Q, and understand the mechanism of this single molecule. Nonetheless, although many existing approaches can remove the noise from the observed data to obtain the estimated mean signal $\hat{\boldsymbol{\mu}}$, it is often impossible to fully reconstruct $x(t)$ due to several practical issues.

First, the minimum time resolution of the experiment limits the information we may learn from the data. Given $\hat{\mu}_i \neq \hat{\mu}_{i+1}$, we can only deduce that at least one state transition has taken place between t_i and t_{i+1}. Worse, when the transitions between states occur much faster than the minimum time resolution, we may fail to detect any transitions taking place between t_i and t_{i+1}. Second, in most single-molecule experiments, parameters (as well as the means) that specify the distribution of measurements can be the same for varying underlying states, and each unique value in $\boldsymbol{\mu}$ can represent more than one discrete state. For instance, in the single-molecule experiment of enzymatic reaction (equation (6)), the fluorescence tag used for tracking the activity of the single enzyme is active at state E^0 but remains dormant at both E and ES. Therefore, although the high values in the observed signal represent state E^0, low values may indicate either state E or ES. Consequently, it is impossible to determine any transitions between E and ES based on the strength of the mean signal alone.

Without knowing the complete trajectory $x(t)$, there is no direct ways of inferring rate parameters in the generator matrix. In addition, it can be more challenging to choose between CTMC models with different numbers of states and transitional structures derived from competing scientific hypotheses. Thus, although scientists have proposed many innovative methods for tracking single-molecule dynamics, sophisticated mathematical and statistical methods are necessary to analyze experimental data and answer relevant questions.

4.2 Model-Free Approaches for Segmenting Single-Molecule Signal

A common step in analyzing the stepwise single-molecule signal is to first remove the noise so that the observed signal can be properly segmented. The distribution of the lengths of the segments that share the same mean, as well as the times and patterns of transitions between segments with different means, can provide useful insight into the mechanism of the single-molecule system. Many conventional statistical tools can be applied for this purpose without the explicit application of the CTMC model. We will call such approaches "model-free" approaches and discuss them in this section.

In model-free approaches, transitions between states are either treated as deterministic or modeled with convenient assumptions, and the major goal is to estimate the mean signal from the noisy data. Straightforward filter-and-thresholding approaches have been used to analyze ion channel recording and the fluorescence intensity signal for a long time (Chung and Kennedy, 1991; VanDongen, 1996; Haran, 2004). In these approaches, a filter algorithm is first applied to reduce the noise in the signal. The transition points can then be determined by establishing a threshold for the changes in the signal strength so that the mean signal can be estimated in a segment-wise fashion. Researchers have also considered other more sophisticated de-noising techniques such as the wavelet in recent years (Taylor, Makarov and Landes, 2010).

In the statistics literature, such questions are known as change-point problems (Chernoff and Zacks, 1964). In a typical change-point model, the entire signal is modeled as a sequence of segments separated by change-points. Within the same segment, the observations are independently and identically distributed. Between successive segments, the observations are drawn independently from the same distributional family with different parameter values (usually with distinctive means). Maximum likelihood estimation can be used to estimate the locations of change-points when their number is known in advance (Hinkley, 1970). Otherwise, the penalized maximum likelihood can be used to select the optimal change-point configuration (Yao, 1988; Braun, Braun and Müller, 2000; Zhang and Siegmund, 2007; Boysen et al., 2009). Because introducing a prior distribution would automatically penalize too many change-points, Bayesian methods are commonly used as well (Smith, 1975; Barry and Hartigan, 1993; Chib, 1998; Fearnhead and Liu, 2007; Du, Kao and Kou, 2016).

Many of the aforementioned change-point methods can be directly applied to a single-molecule signal. Nonetheless, a large number of the change-point algorithms were developed under the equal-variance assumption; that is, observations from different segments share the same common variance, which can be inappropriate in many single-molecule experiments. Moreover, some single-molecule signals contain frequent transitions and short segments, which could be challenging for conventional change-point detection algorithms. In recent years, new methods tailored to single-molecule experiments have been developed. To model the fluctuation in ion channel recording, a Bayesian sampling algorithm designed for detecting changes in the opening probabilities was used (Siekmann, Sneyd and Crampin, 2014). A marginal likelihood approach, where the prior distributions are constructed using an empirical Bayesian procedure, was shown to be effective for signals with nonconstant variance and frequent change-points (Du, Kao and Kou, 2016). A similar Bayesian approach was also used to estimate the number of active fluorescent subunits from photobleaching time traces (Tsekouras et al., 2016). Approaches based on multiscale statistics can be better adapted to the changes in local variation that are common among single-molecule signals (Frick, Munk and Sieling, 2014; Pein, Sieling and Munk, 2017). To handle the scenario where multiple channels were recorded simultaneously, a multivariate change-point method was applied to detect the transitions in multivariate time series using Hotellings T^2 test statistic (Bauer et al., 2018).

4.3 CTMC-Based Approaches for Analyzing Single-Molecule Signal

In contrast to the model-free approaches, model-based approaches rely on a concrete CTMC model to model the state transitions. Within the framework of the CTMC model, the distribution of sojourn time in particular state(s) as well as the transition probabilities between states can be derived or estimated. Although the effectiveness of the CTMC-based approach largely depends on how well the chosen model matches the real single-molecule system, the interpretability of the CTMC model is often worth the risk.

Fitting the dwell time distribution. A straightforward way to learn the underlying CTMC model from the observed single-molecule signal is to analyze the histogram of dwell times in various states. Given a CTMC model with a finite number of states, it is not hard, at least in principle, to derive the analytical formula of the densities of dwell times in a given set of states as functions of the transition rates. Specifically, the density of the dwell time in a single state is exponential, and the density of the dwell time in a particular set of states is a mixture of exponential functions. As long as we can segment the signal properly and match each segment to the underlying discrete state(s), we can obtain the empirical distribution of dwell times and fit the theoretical density function accordingly. This strategy can usually provide direct insight into the mechanisms of a single molecular system, but it is often limited to relatively simple models due to the complexity of the analytical density function.

This method was first used to analyze ion channels recordings, where signals often switch between two levels corresponding to the open and closed state(s) (Colquhoun and Hawkes, 1981). For instance, in model (7), the density of the dwell time in closed states T and AT is a double-exponential function, and the density of the dwell time in open state AR is a single-exponential function. To choose between competing models with different numbers of intermediate states or patterns of state transitions, we can derive the theoretical density function under each model and apply standard model selection metrics, such as the likelihood ratio and AIC (Horn, 1987) or BIC (Ball and Sansom, 1989). Similar strategies were later used to analyze the time-binned fluorescence intensity signal with single-exponential (Zhuang et al., 2000; McKinney et al., 2003) or multiple-exponential functions (Lu, Xun and Xie, 1998; Zhuang et al., 2002; Bokinsky et al., 2003). For the time-stamped photon sequence, the autocorrelation function of waiting times between successive events of photon arrivals is either a single-exponential or a multiple-exponential function and can thus be fitted in a similar fashion (Yang and Xie, 2002; Yang et al., 2003). In recent years, more rigorous research on fitting such multiple-exponential functions has emerged. The roles of the number of observations and the fluctuation in transition rates on fitting the dwell time distribution have been examined (Floyd, Harrison and Van Oijen, 2010). A generalized method of moments was also shown to perform better compared to the traditional least-squares approach (Piatt and Price, 2019).

Hidden Markov model. By treating the observed data as the sum of the mean signal and the stochastic noise whose magnitude may depend on the underlying states, the hidden Markov model (HMM) provides a general way of inferring the CTMC model from single-molecule signals. With the help of the HMM method, the state of the single molecule at any given time of measurement, the mean values of the observed signal, and the transition rates in the generator matrix can be estimated at the same time.

The first systematic work to adopt the HMM to analyze ion channel data appeared in 1990 (Chung et al., 1990). In this work, the Baum–Welch algorithm was used to obtain the maximum likelihood estimators of model parameters and the means of signal. Computational algorithms other than the Baum–Welch algorithm were also explored, such as the direct optimization approach (Qin, Auerbach and Sachs, 2000a) and the segmental k-means method (Qin, 2004). Many researches simply treated the noise in the observed signal as additive white Gaussian noise. However, significant efforts were made to improve the conventional algorithm to handle more realistic model assumptions, especially a signal with non-Gaussian or corrected noise (Venkataramanan, Kuc and Sigworth, 1998; Venkataramanan et al., 1998; Qin, Auerbach and Sachs, 2000b). In addition to the deterministic optimization algorithm, MCMC methods were also popular for inferring parameters and the means of signal (Ball et al., 1999; Rosales et al., 2001; Gin et al., 2009; Siekmann et al., 2011; Epstein et al., 2016).

For fluorescence intensity data, the HMM approaches have been applied to analyze both the time-stamped photon sequence and the time-binned intensity signal. In the time-stamped data, the process of photon arrivals is usually modeled as Poisson process with rates that depend on the underlying states, which allows for the direct application of HMM techniques (Andrec, Levy and Talaga, 2003; Schröder and Grubmüller, 2003; Kou, Xie and Liu, 2005; Okamoto and Sako, 2012; Keller et al., 2014). Notably, the conventional HMM models can be expanded to incorporate the fluctuation in the intensity of photon arrivals due to the diffusion of molecules (Kou, Xie and Liu, 2005). Analysis using time-stamped data can provide a better estimation of the transition times and is more sensitive to rapid transitions compared with analysis using time-binned data. Still, time-binned data are more commonly used in practice due to the added difficulty of collecting and modeling the time-stamped data.

For time-binned data, a maximum likelihood approach was first applied to estimate the means and the transition probabilities in the FRET ratio trajectory under the HMM framework (McKinney, Joo and Ha, 2006). Later, many Bayesian strategies were used to infer parameters in the HMM, such as maximizing the posterior or marginal distribution via the variational Bayes approach (Bronson et al., 2009; Okamoto and Sako, 2012) or empirical Bayes approaches (van de Meent et al., 2014), as well as the MCMC sampling approaches (Chen et al., 2016). Although Gaussian distributional assumptions are often used to model the noise, other distributions, such as Poisson and mixture Gaussian distribution, have also been investigated in the literature as well (Liu et al., 2010). In addition, techniques such as FRET use multiple fluorescence tags to study the dynamic of a single molecule, and the multivariate HMM can be valuable for analyzing multiple channels simultaneously (Liu et al., 2010).

To apply an HMM method, the total number of discrete states in a Markov chain model is needed. From a statistics standpoint, this is a model selection problem. BIC and AIC are often used in practice for the determination of the number of states. However, BIC and AIC encounter conceptual difficulty when the observations $\mathbf{y} = (y(t_1), y(t_2), \dots, y(t_n))$ are supported on the real line with unknown variances, as in the case where $y(t_i)$ follows a Gaussian distribution conditioning on its hidden state. This conceptual difficulty stems from the fact that the HMM can be overly fitted with infinite likelihood as one can make the estimated variance component arbitrarily small (Gassiat and Rousseau, 2014). Marginal likelihood is one method of avoiding the difficulty besetting BIC or AIC and can provide a consistent estimate of the number of states (Chen et al., 2019). Moreover, when multiple trajectories are available for analysis, the number of discrete states estimated from an individual trajectory can vary. If this is the case, a majority rule that reflects the consensus of the observed signals can be used to determine the number of states in the overall model (Chen et al., 2016).

In the study of single-molecule data, it is quite common that the number of distinct means shown in the observed data is smaller than the number of underlying discrete states. Such a scenario will be discussed in the next section. Here, we will stick to the scenario in which each discrete state corresponds to a unique mean.

A straightforward way of determining the number of states is to count the number of modes in the empirical distribution of the signal. We may also first fit a Markov model with many states, and then fine tune the number of states based on the estimated means (Chung et al., 1990; McKinney, Joo and Ha, 2006). Alternatively, we can fit multiple models with different numbers of states and choose the best using the frequentist approaches (McKinney, Joo and Ha, 2006; Liu et al., 2010; Chen et al.,

2016) or Bayesian methods (Bronson et al., 2009; van de Meent et al., 2014; Chen et al., 2019). In recent years, the infinite hidden Markov model (iHMM) has been applied to study single molecule signals (Hines, Bankston and Aldrich, 2015). This approach involves using a hierarchical Dirichlet process as the prior distribution of the transition probability matrix, which guarantees the generation of a proper transition probability matrix with random dimensions. Consequently, iHMM provides a framework for sampling CTMC models with different numbers of states.

4.4 Specific Issues in Analyzing Single-Molecule Signals

The issues discussed in the previous sections can appear in any applications with the HMM. In single-molecule experiments, the application of HMM methods is also beset by unique challenges. In this section, we discuss the problems of handling the aggregation of states, dynamic disorder, and heterogeneity across molecules.

4.4.1 *Resolving aggregated states*. Although determining the number of unique means in the observed signal is a relatively simple matter, specifying the number of discrete states in the CTMC model is more difficult in a single-molecule experiment and deserves special attention. As we have discussed in Section 4.1, although the stochastic model of the molecular dynamic can involve many intermediate states, changes in the means of the observed signal often do not contain sufficient information to discriminate these states. A unique mean in the observed signal may correspond to multiple underlying discrete states. Such phenomena are known as the aggregation of states in the ion channel literature and are also commonly encountered in experiments utilizing fluorescence spectroscopy.

No matter how complicated the true model may be, it is always feasible to first fit the data with a Markov model whose number of states equals to the number of distinctive means in the observed signal. This procedure can still yield a reasonably good segmentation of the observed signal (Fredkin and Rice 1992a, 1992b). We can then investigate the histograms of the lengths of segments with the same means and determine whether we need a more complicated model.

Some HMM-based methods directly fit the observed data with a sophisticated Markov model with aggregated states. After all, the distribution of the lengths of segments and the transition patterns may carry information for distinguishing aggregated states. These algorithms often add constraints to the conventional HMM algorithm to ensure that the aggregated states share the same mean (Rosales, 2004). Techniques such as the reversible-jump MCMC can be used to choose between models with different degrees of aggregations (Hodgson and Green, 1999; de Gunst and Schouten, 2003). Recently, the application of the iHMM has also demonstrated that the aggregated states can be resolved if the dwell times in different aggregated states have distinctive characteristics (Hines, Bankston and Aldrich, 2015).

Nonetheless, the identifiability issue may arise when the observed data do not contain sufficient information to resolve the aggregated states. Many studies have demonstrated that, for some Markov models with aggregated states, the estimations of transition rates can be highly inaccurate or subject to slow convergence (Fredkin and Rice, 1992b; Ball et al., 1999; Hodgson and Green, 1999; Qin, Auerbach and Sachs, 2000a). It remains unknown whether a formal procedure exists to determine whether a given model is identifiable. However, some have suggested that well-designed MCMC sampling algorithms can provide a clue to the nonidentifiability of model parameters based on the pattern of posterior distribution and the speed of convergence (Gin et al., 2009; Siekmann et al., 2011; Siekmann, Sneyd and Crampin, 2014).

4.4.2 *Dynamic disorder.* The analysis of the fluorescence signal from single molecules has confirmed one important feature of single-molecule kinetics: dynamic disorder (Zwanzig, 1992). Simply put, the kinetic rate parameters of molecules, rather than being constant, may fluctuate over time. In a CTMC model, fluctuation in rate parameters can induce a memory effect in the otherwise memoryless Markov process. Taking the enzymatic reaction in (6) as an example, if we define a single turnover as the time for an enzyme to complete one catalytic cycle—that is, the first passage time from state E to state E^0—we can expect successive turnover times to be independent. However, if the rate parameters can change over time, a memory effect—nonzero correlation between successive turnover times—may appear, as has been observed in single-molecule experiments (Lu, Xun and Xie, 1998; Min et al., 2005; Kou, 2008b).

Two ways of modeling dynamic disorder exist: the discrete model and the continuous model. In the discrete model, the molecule can switch between different conformations, each of which comes with a distinct set of rate parameters (Schenter, Lu and Xie, 1999; Berezhkovskii, Szabo and Weiss, 2000; Yang and Cao, 2001; Kou et al., 2005; Chung and Gopich, 2014; Chung, Louis and Gopich, 2016). Such a model effectively expands the state space of the original CTMC model. For instance, by replacing each of the original three states with n conformational states, the enzymatic reaction model (6) can be expanded into the following multiconformational-state model (Kou et al., 2005; English et al., 2006; Du and Kou, 2012):

$$
\begin{array}{ccccccc}
S + E_1 \underset{k_{-11}}{\overset{k_{11}[S]}{\rightleftharpoons}} & ES_1 & \xrightarrow{k_{21}} & P + E_1^0 & & E_1^0 \xrightarrow{\delta_1} E_1 \\
\downarrow\uparrow & \downarrow\uparrow & & \downarrow\uparrow & & \cdots \\
S + E_2 \underset{k_{-12}}{\overset{k_{12}[S]}{\rightleftharpoons}} & ES_2 & \xrightarrow{k_{22}} & P + E_2^0 & & E_2^0 \xrightarrow{\delta_2} E_2 \\
\cdots & \cdots & & \cdots & & \cdots \\
\downarrow\uparrow & \downarrow\uparrow & & \downarrow\uparrow & & \\
S + E_n \underset{k_{-1n}}{\overset{k_{1n}[S]}{\rightleftharpoons}} & ES_n & \xrightarrow{k_{2n}} & P + E_n^0 & & E_n^0 \xrightarrow{\delta_n} E_n
\end{array}
\tag{8}
$$

The multiconformational-state models are capable of modeling dynamic disorder within a Markov framework. Nonetheless, the additional states also increase the degree of state aggregation. Consequently, although conventional HMM-based approaches can technically be used to analyze these models (Chung and Gopich, 2014; Chung, Louis and Gopich, 2016), parameters in multiconformational-state models are usually estimated by fitting analytical functions such as the density function or the autocorrelation function of experimental data. Although these functions are the sums of multiple-exponential decays in principle, explicit analytical formulas are generally available only for small n. Therefore, approximated approaches, such as treating rate parameters in different conformations states as random variables or focusing on the dominated exponential components, are often needed (Kou et al., 2005; Du and Kou, 2012).

The continuous model treats the rate parameters in the CTMC as stochastic processes. Such a model does not introduce additional discrete states but rather sacrifices the Markov property. For instance, the following model was used to study the forming and breaking of intramolecular pairing in a DNA hairpin (Kou, Xie and Liu, 2005):

$$
A \underset{k_{21}\exp[-x(t)]}{\overset{k_{12}\exp[-x(t)]}{\rightleftharpoons}} B, \quad dx_t = -\lambda x_t\, dt + \sqrt{2\xi\lambda}\, dW_t, \tag{9}
$$

where the transition rates between states A and B are modeled with an Ornstein–Uhlenbeck process representing the diffusion of the molecule. Parameters in this model can be inferred from the posterior samples drawn with the help of the data augmentation and MCMC methods.

Both of the aforementioned models handle dynamic disorder by introducing an added layer of complexity that reflects the heterogeneity over time in single-molecule systems. Although the presence of such heterogeneity can be detected from simple statistics, such as the autocorrelation function, how to best model this phenomenon remains unclear. Although complex models may better resemble the actual molecular dynamics, the observed data may not contain sufficient information to reconstruct such subtle dynamics. Some authors have demonstrated that in certain scenarios, complex multi-conformational-state models or continuous models may not necessarily offer significant improvement over a simple two-by-two model where the molecule only switches between two conformational states in each stage (e.g., $n = 2$ in equation (8); Lu, Xun and Xie, 1998; Schenter, Lu and Xie, 1999).

In this regard, nonparametric approaches can be a valuable tool to analyze heterogeneity over time in the experimental data. For instance, the time-stamped photon arrival sequence can be modeled as a doubly stochastic Poisson process in which the arrival rate is also a stochastic process. Then, the arrival rate over time along with the autocorrelation function can be estimated directly using a nonparametric approach without explicit assumptions (Zhang and Kou, 2010). Such model-free estimation may offer valuable information on the extent of heterogeneity as well as provide directions to construct better parametric models.

4.4.3 *Heterogeneity across molecules.* Another issue in analyzing single-molecule data is how to handle multiple signals originating from different molecules. In practice, repeated experiments are often conducted, and scientists wish to combine information from all signals. As a result of the heterogeneity between molecules, not only can the transition rates between states differ across molecules but also the means of signals in the same state can vary. Moreover, the space of discrete states can differ between molecules, because a single molecule may not be able to visit all of the discrete states within a single recording.

A simple strategy of handling multiple signals is to treat different signals as independent realizations of one global model. To infer the global model, we can analyze each observed signal first and then pool the results to estimate global parameters (McKinney, Joo and Ha, 2006; Bronson et al., 2009). As the estimated means of the same state would naturally vary between different signals, suitable algorithms are needed to categorize the segments from the trajectory-wise analysis and match them to the states in the global model. The so-called transition density plot (McKinney, Joo and Ha, 2006), a two-dimensional histogram that represents the distribution of signal levels before and after each transition, can be used for this purpose as well as for determining the number of discrete states in the global model. Then, the global transition probabilities at the log scale can be estimated as the averages of the logarithm of transition probabilities inferred from individual signals. Alternatively, because the signals are independent, it is also feasible to apply the HMM to infer the global model directly using all of the signals simultaneously (Blanco and Walter, 2010; Liu et al., 2010). These strategies, although they can be used to pool information for inferring a global model, simply treat the difference between molecules as noise and may overlook heterogeneity between signals.

Various other approaches have been developed to recognize heterogeneity between molecules. In single-molecule cluster analysis (Blanco et al., 2015) for FRET data, each signal is first segmented using HMM independently. Every estimated segment is then matched to one of the ten global states with the given FRET values. Finally, a modified k-means algorithm is used to cluster signals based on the similarity in the transition probability matrices. This approach enforces a degree of consistency over the signal means but allows the pattern of state transitions to differ. In the method known as the single-molecule analysis of complex kinetic sequences (Schmid and Hugel, 2018), a global CTMC model is used to model the transitions between hidden states, whereas the levels of the observed signals are allowed to differ across molecules.

The hierarchical HMM model can systematically examine the common traits as well as the specific characteristics of individual signals (van de Meent et al., 2014; Chen et al., 2016). In a typical hierarchical HMM setup, all signals share the same global state space, while the distributional parameters, such as the signal means, are drawn from a common prior distribution. Similarly, the transition rate matrices can be assumed to be either constants or independent samples from a prior distribution. Moreover, as a result of either the molecular heterogeneity or the limit duration of the observed signal, a molecule may only visit only a part of the global state space during its course. This issue can be handled by introducing a random indicator vector that specifies the subspace that the corresponding molecule visits (Chen et al., 2016). Through sharing information effectively, the hierarchical HMM model could significantly reduce the uncertainties in estimating model parameters (Chen et al., 2016).

5. CONCLUSION

The development of single-molecule experiments has allowed scientists to study the detailed dynamics of individual molecules. As scientists zoom in on the microscopic world, the intrinsic stochasticity of molecular systems emerges as a dominating factor. Unlike the experiments conducted on the macroscopic scale, advanced single-molecule experiments often have to rely on indirect means of collecting the much-needed information for probing the underlying molecular system. At least in the near future, our measurements of single-molecule systems will remain incomplete, and a large portion of such systems will continue to be hidden from our sight. All of these factors have called for extensive applications of statistical methodology in analyzing single-molecule experimental data. Only by applying appropriate statistical methods can we hope to learn useful information from the noisy data and to bridge the gap between the observed information and the hidden mechanisms. It is, then, no wonder that in recent decades, there has been an increasing trend of applying complex statistical methods for analyzing single-molecule data. It is safe to expect that such momentum will continue as new technological advancements and experimental techniques are developed.

The application of statistical methodology in single-molecule experiments also brings many challenges. First, complex mathematical models are used extensively in single-molecule study. These models are often developed from established physical principles and can provide a clear interpretation of the mechanisms of the underlying system (see Qian and Kou, 2014). Many single-molecule experiments are also designed to validate or improve such models. In this regard, successful applications of the statistical method needs to take such models into account. Although some models (such as the CTMC) can fit into the existing inference framework, many others are too complicated for conventional tools. In particular, when the likelihood function is unavailable, inference with single-molecule data often has to rely on the fitting of relatively simple analytical functions. Such a strategy utilizes only a few moments of the data and it remains to be seen whether general approaches can be developed to handle such a scenarios without losing valuable information.

Second, in many studies of single-molecule data, one must choose between competing stochastic models. Although the current statistical literature contains many tools for model selection, these tools are often developed with relatively generic linear models in mind and thus may not always be suitable for choosing between complicated nonlinear stochastic models. Even though the Bayesian approach might provide a general solution for this matter, it can be challenging to design a suitable sampling algorithm that can incorporate competing models under the same framework. In addition, due to the complexity of the model and the incomplete nature of the data, it is often hard to determine whether the available data contain sufficient information for identifying the given models. Such an identification issue has severely affected the ability to use sophisticated and realistic models in analysis, and rigorous statistical methodology is needed to resolve this matter.

Third, modern single-molecule experiments involve the use of sophisticated measurement techniques, which often introduce additional layers of complexity. Complicated algorithms are usually applied to pre-process the raw data with the aim of removing the unwanted effects stemming from the measurement methods. However, in the realm of single-molecule experiments, due to the strong interactions between the measurement methods and the dynamics of molecules, such a procedure may lead to the loss of valuable information. Therefore, for the in-depth application of statistical methodology, researchers needs to consider the data generation process and should attempt to incorporate the relevant information into the inference framework. Such works would require not only a good understanding of the experimental procedure, but also the ability to design a new inference approach to accommodate the extended model and data set.

With these issues in mind, we hope that our review of statistical methodologies in single-molecule experiments will not only provide a general picture of the current development in this area but also ignite further interest in introducing and developing new methodologies for analyzing single-molecule data.

ACKNOWLEDGEMENT

The authors are grateful to Professor David Siegmund for very constructive suggestions and comments. S. C. Kou's research is supported in part by NSF Grant DMS-1810914.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical Methodology in Single-Molecule Experiments" (DOI: 10.1214/19-STS752SUPP; .pdf). Supplementary information, including more discussion of the history of single-molecule experiments and more discussion of the motion of a single molecule and models beyond diffusion.

REFERENCES

ANDREC, M., LEVY, R. M. and TALAGA, D. S. (2003). Direct determination of kinetic rates from single-molecule photon arrival trajectories using hidden Markov models. *J. Phys. Chem. A* **107** 7454–7464. https://doi.org/10.1021/jp035514+

ARCIZET, D., MEIER, B., SACKMANN, E., RÄDLER, J. O. and HEINRICH, D. (2008). Temporal analysis of active and passive transport in living cells. *Phys. Rev. Lett.* **101** 248103. https://doi.org/10.1103/PhysRevLett.101.248103

BALL, F. G. and SANSOM, M. S. P. (1989). Ion-channel gating mechanisms: Model identification and parameter estimation from single channel recordings. *Proc. R. Soc. Lond., B Biol. Sci.* **236** 385–416.

BALL, F. G., CAI, Y., KADANE, J. B. and O'HAGAN, A. (1999). Bayesian inference for ion-channel gating mechanisms directly from single-channel recordings, using Markov chain Monte Carlo. *Proc. R. Soc. Lond. A Mat.* **455** 2879–2932.

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

BAUER, M., LI, C., MÜLLEN, K., BASCHÉ, T. and HINZE, G. (2018). State transition identification in multivariate time series (STIMTS) applied to rotational jump trajectories from single molecules. *J. Chem. Phys.* **149** 164104.

BEREZHKOVSKII, A. M., SZABO, A. and WEISS, G. H. (2000). Theory of the fluorescence of single molecules undergoing multistate conformational dynamics. *J. Phys. Chem. B* **104** 3776–3780.

BERGLUND, A. J. (2010). Statistics of camera-based single-particle tracking. *Phys. Rev. E* **82** 011917.

BERNSTEIN, J. and FRICKS, J. (2016). Analysis of single particle diffusion with transient binding using particle filtering. *J. Theoret. Biol.* **401** 109–121.

BLAINEY, P. C., LUO, G., KOU, S. C., MANGEL, W. F., VERDINE, G. L., BAGCHI, B. and XIE, S. X. (2009). Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16** 1224–1229.

BLANCO, M. and WALTER, N. G. (2010). Analysis of complex single-molecule FRET time trajectories. In *Methods in Enzymology* **472** 153–178. Elsevier, Amsterdam.

BLANCO, M. R., MARTIN, J. S., KAHLSCHEUER, M. L., KRISHNAN, R., ABELSON, J., LAEDERACH, A. and WALTER, N. G. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics. *Nat. Methods* **12** 1077–1084. https://doi.org/10.1038/nmeth.3602

BOKINSKY, G., RUEDA, D., MISRA, V. K., RHODES, M. M., GORDUS, A., BABCOCK, H. P., WALTER, N. G. and ZHUANG, X. (2003). Single-molecule transition-state analysis of RNA folding. *Proc. Natl. Acad. Sci. USA* **100** 9302–9307.

BOSCH, P. J., KANGER, J. S. and SUBRAMANIAM, V. (2014). Classification of dynamical diffusion states in single molecule tracking microscopy. *Biophys. J.* **107** 588–598. https://doi.org/10.1016/j.bpj.2014.05.049

BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348 https://doi.org/10.1214/07-AOS558

BRAUN, J. V., BRAUN, R. K. and MÜLLER, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314. MR1782480 https://doi.org/10.1093/biomet/87.2.301

BROCK, R., HINK, M. A. and JOVIN, T. M. (1998). Fluorescence correlation microscopy of cells in the presence of autofluorescence. *Biophys. J.* **75** 2547–2557.

BRONSON, J. E., FEI, J., HOFMAN, J. M., GONZALEZ JR, R. L. and WIGGINS, C. H. (2009). Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* **97** 3196–3205.

CHEN, Y., FUH, C.-D., KAO, C.-L. and KOU, S. C. (2019). Determine the number of states in hidden Markov models via marginal likelihood. *Preprint*.

CHEN, Y., MÜLLER, J. D., SO, P. T. C. and GRATTON, E. (1999). The photon counting histogram in fluorescence fluctuation spectroscopy. *Biophys. J.* **77** 553–567.

CHEN, Y., SHEN, K., SHAN, S.-O. and KOU, S. C. (2016). Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *J. Amer. Statist. Assoc.* **111** 951–966. MR3561922 https://doi.org/10.1080/01621459.2016.1140050

CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Stat.* **35** 999–1018. MR0179874 https://doi.org/10.1214/aoms/1177700517

CHIB, S. (1998). Estimation and comparison of multiple changepoint models. *J. Econometrics* **86** 221–241. MR1649222 https://doi.org/10.1016/S0304-4076(97)00115-2

CHUNG, H. S. and GOPICH, I. V. (2014). Fast single-molecule FRET spectroscopy: Theory and experiment. *Phys. Chem. Chem. Phys.* **16** 18644–18657. https://doi.org/10.1039/c4cp02489c

CHUNG, S. H. and KENNEDY, R. A. (1991). Forward-backward non-linear filtering technique for extracting small biological signals from noise. *J. Neurosci. Methods* **40** 71–86. https://doi.org/10.1016/0165-0270(91)90118-j

CHUNG, H. S., LOUIS, J. M. and GOPICH, I. V. (2016). Analysis of fluorescence lifetime and energy transfer efficiency in single-molecule photon trajectories of fast-folding proteins. *J. Phys. Chem. B* **120** 680–699.

CHUNG, S.-H., MOORE, J. B., XIA, L., PREMKUMAR, L. S. and GAGE, P. W. (1990). Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **329** 265–285.

CHUNG, I., AKITA, R., VANDLEN, R., TOOMRE, D., SCHLESSINGER, J. and MELLMAN, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells. *Nature* **464** 783–787.

CLAUSEN, M. P. and LAGERHOLM, C. B. (2013). Visualization of plasma membrane compartmentalization by high-speed quantum dot tracking. *Nano Lett.* **13** 2332–2337.

COLQUHOUN, D. and HAWKES, A. G. (1981). On the stochastic properties of single ion channels. *Proc. R. Soc. Lond., B Biol. Sci.* **211** 205–235.

DAS, R., CAIRO, C. W. and COOMBS, D. (2009). A hidden Markov model for single particle tracks quantifies dynamic interactions between LFA-1 and the actin cytoskeleton. *PLoS Comput. Biol.* **5** e1000556, 16. MR2577427 https://doi.org/10.1371/journal.pcbi.1000556

DE GUNST, M. C. M. and SCHOUTEN, B. (2003). Model selection for hidden Markov models of ion channel data by reversible jump Markov chain Monte Carlo. *Bernoulli* **9** 373–393. MR1997489 https://doi.org/10.3150/bj/1065444810

DEL CASTILLO, J. and KATZ, B. (1957). Interaction at end-plate receptors between different choline derivatives. *Proc. R. Soc. Lond., B Biol. Sci.* **146** 369–381.

DU, C., KAO, C.-L. M. and KOU, S. C. (2016). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc.* **111** 314–330. MR3494662 https://doi.org/10.1080/01621459.2015.1006365

DU, C. and KOU, S. C. (2012). Correlation analysis of enzymatic reaction of a single protein molecule. *Ann. Appl. Stat.* **6** 950–976. MR3012516 https://doi.org/10.1214/12-AOAS541

DU, C. and KOU, S. C. (2020). Supplement to "Statistical methodology in single-molecule experiments." https://doi.org/10.1214/19-STS752SUPP.

ELSON, E. L. (2011). Fluorescence correlation spectroscopy: Past, present, future. *Biophys. J.* **101** 2855–2870. https://doi.org/10.1016/j.bpj.2011.11.012

ELSON, E. L. and MAGDE, D. (1974). Fluorescence correlation spectroscopy. I. Conceptual basis and theory. *Biopolymers: Original Research on Biomolecules* **13** 1–27.

ENDERLEIN, J., GREGOR, I., PATRA, D. and FITTER, J. (2005). Statistical analysis of diffusion coefficient determination by fluorescence correlation spectroscopy. *J. Fluoresc.* **15** 415–422.

ENGLISH, B. P., MIN, W., VAN OIJEN, A. M., LEE, K. T., LUO, G., SUN, H., CHERAYIL, B. J., KOU, S. C. and XIE, X. S. (2006). Ever-fluctuating single enzyme molecules: Michaelis–Menten equation revisited. *Nat. Chem. Biol.* **2** 87–94.

EPSTEIN, M., CALDERHEAD, B., GIROLAMI, M. A. and SIVILOTTI, L. G. (2016). Bayesian statistical inference in ion-channel models with exact missed event correction. *Biophys. J.* **111** 333–348.

FEARNHEAD, P. and LIU, Z. (2007). On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 589–605. MR2370070 https://doi.org/10.1111/j.1467-9868.2007.00601.x

FLOYD, D. L., HARRISON, S. C. and VAN OIJEN, A. M. (2010). Analysis of kinetic intermediates in single-particle dwell-time distributions. *Biophys. J.* **99** 360–366.

FREDKIN, D. R. and RICE, J. A. (1992a). Bayesian restoration of single-channel patch clamp recordings. *Biometrics* 427–448.

FREDKIN, D. R. and RICE, J. A. (1992b). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. R. Soc. Lond., B Biol. Sci.* **249** 125–132.

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

GASSIAT, E. and ROUSSEAU, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli* **20** 2039–2075. MR3263098 https://doi.org/10.3150/13-BEJ550

GENNERICH, A. and SCHILD, D. (2000). Fluorescence correlation spectroscopy in small cytosolic compartments depends critically on the diffusion model used. *Biophys. J.* **79** 3294–3306.

GIN, E., FALCKE, M., WAGNER, L. E., YULE, D. I. and SNEYD, J. (2009). Markov chain Monte Carlo fitting of single-channel data from inositol trisphosphate receptors. *J. Theoret. Biol.* **257** 460–474.

GLOTER, A. and JACOD, J. (2001). Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM Probab. Stat.* **5** 225–242. MR1875672 https://doi.org/10.1051/ps:2001110

HA, T., ENDERLE, T., OGLETREE, D. F., CHEMLA, D. S., SELVIN, P. R. and WEISS, S. (1996). Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. USA* **93** 6264–6268.

HARAN, G. (2004). Noise reduction in single-molecule fluorescence trajectories of folding proteins. *Chem. Phys.* **307** 137–145.

HE, J., GUO, S.-M. and BATHE, M. (2012). Bayesian approach to the analysis of fluorescence correlation spectroscopy data I: Theory. *Anal. Chem.* **84** 3871–3879.

HINES, K. E., BANKSTON, J. R. and ALDRICH, R. W. (2015). Analyzing single-molecule time series via nonparametric Bayesian inference. *Biophys. J.* **108** 540–556.

HINKLEY, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57** 1–17. MR0273727 https://doi.org/10.1093/biomet/57.1.1

HODGSON, M. E. A. and GREEN, P. J. (1999). Bayesian choice among Markov models of ion channels using Markov chain Monte Carlo. *Proc. R. Soc. Lond. A Mat.* **455** 3425–3448.

HORN, R. (1987). Statistical methods for model discrimination. Applications to gating kinetics and permeation of the acetylcholine receptor channel. *Biophys. J.* **51** 255–263.

HUET, S., KARATEKIN, E., TRAN, V. S., FANGET, I., CRIBIER, S. and HENRY, J.-P. (2006). Analysis of transient behavior in complex trajectories: Application to secretory vesicle dynamics. *Biophys. J.* **91** 3542–3559. https://doi.org/10.1529/biophysj.105.080622

JEON, J.-H. and METZLER, R. (2010). Fractional Brownian motion and motion governed by the fractional Langevin equation in confined geometries. *Phys. Rev. E* **81** 021103, 11. MR2610879 https://doi.org/10.1103/PhysRevE.81.021103

KASK, P., PALO, K., FAY, N., BRAND, L., METS, Ü., ULLMANN, D., JUNGMANN, J., PSCHORR, J. and GALL, K. (2000). Two-dimensional fluorescence intensity distribution analysis: Theory and applications. *Biophys. J.* **78** 1703–1713.

KELLER, B. G., KOBITSKI, A., JÄSCHKE, A., NIENHAUS, G. U. and NOÉ, F. (2014). Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *J. Am. Chem. Soc.* **136** 4534–4543.

KEPTEN, E., BRONSHTEIN, I. and GARINI, Y. (2013). Improved estimation of anomalous diffusion exponents in single-particle tracking experiments. *Phys. Rev. E* **87** 052713.

KEPTEN, E., WERON, A., SIKORA, G., BURNECKI, K. and GARINI, Y. (2015). Guidelines for the fitting of anomalous diffusion mean square displacement graphs from single particle tracking experiments. *PLoS ONE* **10** e0117722. https://doi.org/10.1371/journal.pone.0117722

KOO, P. K. and MOCHRIE, S. G. J. (2016). Systems-level approach to uncovering diffusive states and their transitions from single-particle trajectories. *Phys. Rev. E* **94** 052412. https://doi.org/10.1103/PhysRevE.94.052412

KOO, P. K., WEITZMAN, M., SABANAYGAM, C. R., VAN GOLEN, K. L. and MOCHRIE, S. G. J. (2015). Extracting diffusive states of Rho GTPase in live cells: Towards in vivo biochemistry. *PLoS Comput. Biol.* **11** e1004297.

KOPPEL, D. E. (1974). Statistical accuracy in fluorescence correlation spectroscopy. *Phys. Rev. A* **10** 1938.

KOU, S. C. (2008a). Stochastic modeling in nanoscale biophysics: Subdiffusion within proteins. *Ann. Appl. Stat.* **2** 501–535. MR2524344 https://doi.org/10.1214/07-AOAS149

KOU, S. C. (2008b). Stochastic networks in nanoscale biophysics: Modeling enzymatic reaction of a single protein. *J. Amer. Statist. Assoc.* **103** 961–975. MR2462886 https://doi.org/10.1198/016214507000001021

KOU, S. C. and XIE, X. S. (2004). Generalized Langevin equation with fractional Gaussian noise: Subdiffusion within a single protein molecule. *Phys. Rev. Lett.* **93** 180603.

KOU, S. C., XIE, X. S. and LIU, J. S. (2005). Bayesian analysis of single-molecule experimental data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 469–506. MR2137252 https://doi.org/10.1111/j.1467-9876.2005.00509.x

KOU, S. C., CHERAYIL, B. J., MIN, W., ENGLISH, B. P. and XIE, X. S. (2005). Single-molecule Michaelis–Menten equations. *J. Phys. Chem. B* **109** 19068–19081.

KUSUMI, A., SAKO, Y. and YAMAMOTO, M. (1993). Confined lateral diffusion of membrane receptors as studied by single particle tracking (nanovid microscopy). Effects of calcium-induced differentiation in cultured epithelial cells. *Biophys. J.* **65** 2021–2040.

LIU, Y., PARK, J., DAHMEN, K. A., CHEMLA, Y. R. and HA, T. (2010). A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B* **114** 5386–5403.

LU, H. P., XUN, L. and XIE, X. S. (1998). Single-molecule enzymatic dynamics. *Science* **282** 1877–1882.

MAGDE, D., ELSON, E. and WEBB, W. W. (1972). Thermodynamic fluctuations in a reacting systemmeasurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.* **29** 705–708.

MCKINNEY, S. A., JOO, C. and HA, T. (2006). Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* **91** 1941–1951.

MCKINNEY, S. A., DÉCLAIS, A.-C., LILLEY, D. M. J. and HA, T. (2003). Structural dynamics of individual holliday junctions. *Nat. Struct. Mol. Biol.* **10** 93–97.

MEILHAC, N., LE GUYADER, L., SALOME, L. and DESTAINVILLE, N. (2006). Detection of confinement and jumps in single-molecule membrane trajectories. *Phys. Rev. E* **73** 011915.

MELNYKOV, A. V. and HALL, K. B. (2009). Revival of high-order fluorescence correlation analysis: Generalized theory and biochemical applications. *J. Phys. Chem. B* **113** 15629–15638. https://doi.org/10.1021/jp906539k

MESETH, U., WOHLAND, T., RIGLER, R. and VOGEL, H. (1999). Resolution of fluorescence correlation measurements. *Biophys. J.* **76** 1619–1631.

METZLER, R. and KLAFTER, J. (2000). The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Phys. Rep.* **339** 77. MR1809268 https://doi.org/10.1016/S0370-1573(00)00070-3

METZLER, R., JEON, J.-H., CHERSTVY, A. G. and BARKAI, E. (2014). Anomalous diffusion models and their properties: Non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* **16** 24128–24164.

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, 13. MR2788037 https://doi.org/10.1103/PhysRevE.82.041914

MIN, W., ENGLISH, B., LUO, G., CHERAYIL, B., KOU, S. C. and XIE, X. S. (2005). Fluctuating enzymes: Lessons from single-molecule studies. *Acc. Chem. Res.* **38** 923–931.

MONNIER, N., GUO, S.-M., MORI, M., HE, J., LÉNÁRT, P. and BATHE, M. (2012). Bayesian approach to MSD-based analysis of particle motion in live cells. *Biophys. J.* **103** 616–626.

MONNIER, N., BARRY, Z., PARK, H. Y., SU, K.-C., KATZ, Z., ENGLISH, B. P., DEY, A., PAN, K., CHEESEMAN, I. M. et al. (2015). Inferring transient particle transport dynamics in live cells. *Nat. Methods* **12** 838–840.

MÜLLER, J. D. (2004). Cumulant analysis in fluorescence fluctuation spectroscopy. *Biophys. J.* **86** 3981–3992.

MÜLLER, J. D., CHEN, Y. and GRATTON, E. (2000). Resolving heterogeneity on the single molecular level with the photon-counting histogram. *Biophys. J.* **78** 474–486.

NEHER, E. and SAKMANN, B. (1976). Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature* **260** 799–802. https://doi.org/10.1038/260799a0

OKAMOTO, K. and SAKO, Y. (2012). Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.* **103** 1315–1324.

ORRIT, M., HA, T. and SANDOGHDAR, V. (2014). Single-molecule optical spectroscopy. *Chem. Soc. Rev.* **43** 973–976.

OTT, M., SHAI, Y. and HARAN, G. (2013). Single-particle tracking reveals switching of the HIV fusion peptide between two diffusive modes in membranes. *J. Phys. Chem. B* **117** 13308–13321. https://doi.org/10.1021/jp4039418

PEIN, F., SIELING, H. and MUNK, A. (2017). Heterogeneous change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1207–1227. MR3689315 https://doi.org/10.1111/rssb.12202

PERSSON, F., LINDÉN, M., UNOSON, C. and ELF, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods* **10** 265–269. https://doi.org/10.1038/nmeth.2367

PIATT, S. and PRICE, A. C. (2019). Analyzing dwell times with the generalized method of moments. *PLoS ONE* **14** e0197726. https://doi.org/10.1371/journal.pone.0197726

QIAN, H. and ELSON, E. L. (1990a). Distribution of molecular aggregation by analysis of fluctuation moments. *Proc. Natl. Acad. Sci. USA* **87** 5479–5483.

QIAN, H. and ELSON, E. L. (1990b). On the analysis of high order moments of fluorescence fluctuations. *Biophys. J.* **57** 375–380.

QIAN, H. and KOU, S. C. (2014). Statistics and related topics in single-molecule biophysics. *Annu. Rev. Stat. Appl.* **1** 465–492. https://doi.org/10.1146/annurev-statistics-022513-115535

QIAN, H., SHEETZ, M. P. and ELSON, E. L. (1991). Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. *Biophys. J.* **60** 910–921.

QIN, F. (2004). Restoration of single-channel currents using the segmental k-means method based on hidden Markov modeling. *Biophys. J.* **86** 1488–1501.

QIN, F., AUERBACH, A. and SACHS, F. (2000a). A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* **79** 1915–1927.

QIN, F., AUERBACH, A. and SACHS, F. (2000b). Hidden Markov modeling for single channel kinetics with filtering and correlated noise. *Biophys*. *J*. **79** 1928–1944.

RENNER, M., WANG, L., LEVI, S., HENNEKINNE, L. and TRILLER, A. (2017). A simple and powerful analysis of lateral subdiffusion using single particle tracking. *Biophys*. *J*. **113** 2452–2463.

ROSALES, R. A. (2004). MCMC for hidden Markov models incorporating aggregation of states and filtering. *Bull*. *Math*. *Biol*. **66** 1173–1199. MR2253818 https://doi.org/10.1016/j.bulm.2003.12.001

ROSALES, R., STARK, J. A., FITZGERALD, W. J. and HLADKY, S. B. (2001). Bayesian restoration of ion channel records using hidden Markov models. *Biophys*. *J*. **80** 1088–1103.

SAKMANN, B. (2013). *Single-Channel Recording*. Springer, Berlin.

SAXTON, M. J. (1993). Lateral diffusion in an archipelago. Single-particle diffusion. *Biophys*. *J*. **64** 1766–1780.

SAXTON, M. J. (1994). Single-particle tracking: Models of directed transport. *Biophys*. *J*. **67** 2110–2119.

SAXTON, M. J. (1997). Single-particle tracking: The distribution of diffusion coefficients. *Biophys*. *J*. **72** 1744–1753.

SAXTON, M. J. and JACOBSON, K. (1997). Single-particle tracking: Applications to membrane dynamics. *Annu*. *Rev*. *Biophys*. *Biomol*. *Struct*. **26** 373–399.

SBALZARINI, I. F. and KOUMOUTSAKOS, P. (2005). Feature point tracking and trajectory analysis for video imaging in cell biology. *J*. *Struct*. *Biol*. **151** 182–195.

SCHENTER, G. K., LU, H. P. and XIE, X. S. (1999). Statistical analyses and theoretical models of single-molecule enzymatic dynamics. *J*. *Phys*. *Chem*. *A* **103** 10477–10488.

SCHMID, S. and HUGEL, T. (2018). Efficient use of single molecule time traces to resolve kinetic rates, models and uncertainties. *J*. *Chem*. *Phys*. **148** 123312. https://doi.org/10.1063/1.5006604

SCHMIDT, T., SCHÜTZ, G. J., BAUMGARTNER, W., GRUBER, H. J. and SCHINDLER, H. (1996). Imaging of single molecule diffusion. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **93** 2926–2929.

SCHRÖDER, G. F. and GRUBMÜLLER, H. (2003). Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments. *J*. *Chem*. *Phys*. **119** 9920–9924.

SIEKMANN, I., SNEYD, J. and CRAMPIN, E. J. (2014). Statistical analysis of modal gating in ion channels. *Proc*. *R*. *Soc*. *A* **470** 20140030.

SIEKMANN, I., WAGNER II, L. E., YULE, D., FOX, C., BRYANT, D., CRAMPIN, E. J. and SNEYD, J. (2011). MCMC estimation of Markov models for ion channels. *Biophys*. *J*. **100** 1919–1929.

SIKORA, G., KEPTEN, E., WERON, A., BALCEREK, M. and BURNECKI, K. (2017a). An efficient algorithm for extracting the magnitude of the measurement error for fractional dynamics. *Phys*. *Chem*. *Chem*. *Phys*. **19** 26566–26581.

SIKORA, G., TEUERLE, M., WYŁOMAŃSKA, A. and GREBENKOV, D. (2017b). Statistical properties of the anomalous scaling exponent estimator based on time-averaged mean-square displacement. *Phys*. *Rev*. *E* **96** 022132. https://doi.org/10.1103/PhysRevE.96.022132

SIMSON, R., SHEETS, E. D. and JACOBSON, K. (1995). Detection of temporary lateral confinement of membrane proteins using single-particle tracking analysis. *Biophys*. *J*. **69** 989–993.

SLATOR, P. J. and BURROUGHS, N. J. (2018). A hidden Markov model for detecting confinement in single-particle tracking trajectories. *Biophys*. *J*. **115** 1741–1754. https://doi.org/10.1016/j.bpj.2018.09.005

SLATOR, P. J., CAIRO, C. W. and BURROUGHS, N. J. (2015). Detection of diffusion heterogeneity in single particle tracking trajectories using a hidden Markov model with measurement noise propagation. *PLoS ONE* **10** e0140759. https://doi.org/10.1371/journal.pone.0140759

SMITH, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* **62** 407–416. MR0381115 https://doi.org/10.1093/biomet/62.2.407

SUH, J., CHOY, K.-L., LAI, S. K., SUK, J. S., TANG, B. C., PRABHU, S. and HANES, J. (2007). PEGylation of nanoparticles improves their cytoplasmic transport. *Int*. *J*. *Nanomed*. **2** 735–741.

SUN, G., GUO, S.-M., TEH, C., KORZH, V., BATHE, M. and WOHLAND, T. (2015). Bayesian model selection applied to the analysis of fluorescence correlation spectroscopy data of fluorescent proteins in vitro and in vivo. *Anal*. *Chem*. **87** 4326–4333.

TAYLOR, J. N., MAKAROV, D. E. and LANDES, C. F. (2010). Denoising single-molecule FRET trajectories with wavelets and Bayesian inference. *Biophys*. *J*. **98** 164–173.

TSEKOURAS, K., CUSTER, T. C., JASHNSAZ, H., WALTER, N. G. and PRESSÉ, S. (2016). A novel method to accurately locate and count large numbers of steps by photobleaching. *Mol*. *Biol*. *Cell* **27** 3601–3615. https://doi.org/10.1091/mbc.E16-06-0404

VANDONGEN, A. M. (1996). A new algorithm for idealizing single ion channel data containing multiple unknown conductance levels. *Biophys*. *J*. **70** 1303–1315.

VAN DE MEENT, J.-W., BRONSON, J. E., WIGGINS, C. H. and GONZALEZ JR, R. L. (2014). Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophys*. *J*. **106** 1327–1337.

VENKATARAMANAN, L., KUC, R. and SIGWORTH, F. J. (1998). Identification of hidden Markov models for ion channel currents. II. State-dependent excess noise. *IEEE Trans*. *Signal Process*. **46** 1916–1929.

VENKATARAMANAN, L., WALSH, J. L., KUC, R. and SIGWORTH, F. J. (1998). Identification of hidden Markov models for ion channel currents. I. Colored background noise. *IEEE Trans*. *Signal Process*. **46** 1901–1915.

VESTERGAARD, C. L., BLAINEY, P. C. and FLYVBJERG, H. (2014). Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys*. *Rev*. *E* **89** 022726.

WAGNER, T., KROLL, A., HARAMAGATTI, C. R., LIPINSKI, H.-G. and WIEMANN, M. (2017). Classification and segmentation of nanoparticle diffusion trajectories in cellular micro environments. *PLoS ONE* **12** e0170165. https://doi.org/10.1371/journal.pone.0170165

WOHLAND, T., RIGLER, R. and VOGEL, H. (2001). The standard deviation in fluorescence correlation spectroscopy. *Biophys*. *J*. **80** 2987–2999.

WU, B. and MÜLLER, J. D. (2005). Time-integrated fluorescence cumulant analysis in fluorescence fluctuation spectroscopy. *Biophys*. *J*. **89** 2721–2735.

WU, Z., BI, H., PAN, S., MENG, L. and ZHAO, X. S. (2016). Determination of equilibrium constant and relative brightness in fluorescence correlation spectroscopy by considering third-order correlations. *J*. *Phys*. *Chem*. *B* **120** 11674–11682. https://doi.org/10.1021/acs.jpcb.6b07953

YANG, S. and CAO, J. (2001). Two-event echos in single-molecule kinetics: A signature of conformational fluctuations. *J*. *Phys*. *Chem*. *B* **105** 6536–6549.

YANG, H. and XIE, X. S. (2002). Statistical approaches for probing single-molecule dynamics photon-by-photon. *Chem*. *Phys*. **284** 423–437.

YANG, H., LUO, G., KARNCHANAPHANURACH, P., LOUIE, T.-M., RECH, I., COVA, S., XUN, L. and XIE, X. S. (2003). Protein conformational dynamics probed by single-molecule electron transfer. *Science* **302** 262–266.

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist*. *Probab*. *Lett*. **6** 181–189. MR0919373 https://doi.org/10.1016/0167-7152(88)90118-6

Yin, S., Song, N. and Yang, H. (2018). Detection of velocity and diffusion coefficient change points in single-particle trajectories. *Biophys. J.* **115** 217–229.

Zhang, T. and Kou, S. C. (2010). Nonparametric inference of doubly stochastic Poisson process data via the kernel method. *Ann. Appl. Stat.* **4** 1913–1941. MR2829941 https://doi.org/10.1214/10-AOAS352

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32, 309. MR2345571 https://doi.org/10.1111/j.1541-0420.2006.00662.x

Zhuang, X., Bartley, L. E., Babcock, H. P., Russell, R., Ha, T., Herschlag, D. and Chu, S. (2000). A single-molecule study of RNA catalysis and folding. *Science* **288** 2048–2051.

Zhuang, X., Kim, H., Pereira, M. J. B., Babcock, H. P., Walter, N. G. and Chu, S. (2002). Correlating structural dynamics and function in single ribozyme molecules. *Science* **296** 1473–1476.

Zwanzig, R. (1992). Dynamical disorder: Passage through a fluctuating bottleneck. *J. Chem. Phys.* **97** 3587–3589.