JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

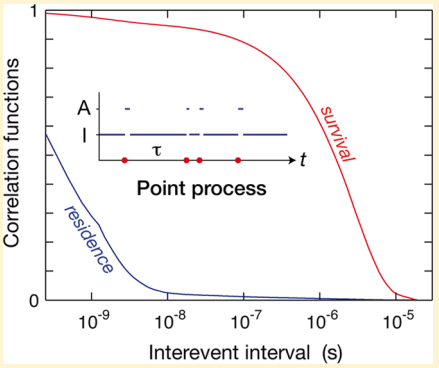
Analysis of Protein Dynamics Simulations by a Stochastic Point Process Approach

Bertil Halle* and Filip Persson

Biophysical Chemistry, Lund University, POB 124, SE-22100 Lund, Sweden

Supporting Information

ABSTRACT: MD simulations can now explore the complex dynamics of proteins and their associated solvent in atomic detail on a millisecond time scale. Among the phenomena that thereby become amenable to detailed study are intermittent conformational transitions where the protein accesses transient high-energy states that often play key roles in biology. Here, we present a coherent theoretical framework, based on the stochastic theory of stationary point processes, that allows the essential dynamical characteristics of such processes to be efficiently extracted from the MD trajectory without assuming Poisson statistics. Since the complete information content of a point process is contained in the sequence of residence or interevent times, the experimentally relevant survival correlation function can be computed several orders of magnitude more efficiently than with the conventional approach, involving averaging over initial times. We also present a detailed analysis of the statistical and binning errors, of particular importance when MD results are compared with experiment. As an illustration of the general theoretical framework, we use a 1 ms MD trajectory of the protein BPTI to analyze the exchange kinetics of an internal water molecule and the dynamics of the rare conformational fluctuations that govern the rate of this exchange process.



1. INTRODUCTION

The rates of protein-mediated elementary processes, such as ligand binding/release, proton exchange, and electron transfer, generally depend on the fluctuating protein conformation. In the simplest case (Figure 1), the protein fluctuates between a 'ground' inactive (I) state, where the elementary process cannot take place, and an 'excited' active (A) state, where the elementary process always occurs before the protein returns to the I state. In this conformational gating limit, the protein typically spends long periods in the I state, intermittently interrupted by brief visits to the A state, and the observed rate of the elementary process equals the I → A transition rate. This scenario was first discussed in connection with amide hydrogen exchange, where it is known as the EX1 limit,[1,2] but similar consideration applies to, for example, ligand binding and release.[3−6]

The ubiquitous dynamical coupling between elementary process and conformational fluctuations provides a major rationale for studying protein dynamics *per se* and, at the same time, furnishes a means for doing so: any experimentally accessible conformation-dependent elementary process, regardless of its biological relevance, can be used to probe the underlying conformational dynamics via a time correlation function that is closely related to the probability that the elementary process has not yet taken place. We refer to this time-dependent survival probability as the survival correlation function (SCF). Our main objective here is to present a new and highly efficient approach for computing the SCF and related quantities from molecular dynamics (MD) simulation data. We also analyze the nonexponential decay of the SCF resulting from dynamical disorder,[5] that is, when the rate of the elementary process is itself a random function of time, for example, because the protein fluctutates among conformational states with different intrinsic rates of the elementary process.

To illustrate the fairly general computational approach presented here, we apply it to the conformationally gated exchange of a buried water molecule in the protein bovine pancreatic trypsin inhibitor (BPTI). The microsecond kinetics of this elementary process have been studied experimentally by ^{2}H magnetic relaxation dispersion (MRD) measurements on rotationally immobilized proteins,[7] which essentially yields the Fourier transform of the SCF.[8] Our computational analysis is based on the 1-ms MD trajectory of hydrated BPTI generated

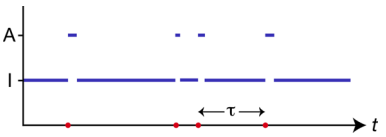


Figure 1. Intermittent two-state transitions represented as a point process along the time (t) axis. The blue line shows transitions between the inactive (I) and active (A) states in the continuous-time trajectory, and the red points represent the transition events. One residence (interevent) time is indicated (τ).

Received: March 1, 2013

   dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

by Shaw et al., who also presented a crude analysis of internal-water exchange.[9]

We are not aware of any previous simulation-based computation of the SCF for internal-water exchange. On the other hand, numerous MD studies have examined water exchange kinetics in the external hydration layer of proteins.[10−19] In these studies, the SCF was computed with an algorithm first used in the context of hydration dynamics by Impey, Madden, and McDonald.[20] However, for intermittent dynamics, the novel algorithm presented here is several orders of magnitude more efficient. Remarkably, none of these studies provided estimates of the statistical uncertainty in the SCF or in the mean survival time (the time integral of the SCF). The present work therefore includes a detailed analysis of both the statistical error and the binning error in the SCF and related quantities.

A deterministic Newtonian trajectory generated by a classical MD simulation of an equilibrium system can often be fruitfully modeled in terms of a stationary stochastic process. In many cases, the conformational fluctuations of a protein can be accurately modeled as stochastic transitions among a small set of discrete states. If the actual transition is fast compared to the typical time interval between transitions, the fluctuations can be modeled as a continuous-time process with discrete state space. In the mathematical literature, such a stochastic process is known as a stationary point process.[21−23] The Poisson process is a familiar special case of the stationary point process.

In the point process formalism, the I → A transitions are described as points on a continuous time axis (Figure 1). The mathematical treatment is most conveniently developed in terms of the time intervals τ between adjacent points (or transitions). We refer to these intervals as residence times (RTs). The series of RTs extracted from the MD trajectory constitutes the complete information content of the process, from which everything else is derived. In particular, we compute the residence correlation function (RCF). This quantity is not experimentally accessible, but it is conceptually and computationally useful. We thus use the RCF as a stepping stone for obtaining the SCF by making use of an exact relationship between these two correlation functions.

The computational approach described here is applicable to a wide range of processes and systems. Ignoring vibrational/librational degrees of freedom, the conformational states accessible to a protein can be described in terms of backbone and side-chains dihedral angles. At least in folded proteins, these angles are confined to narrow ranges and the transitions among different local energy minima (rotamers) tend to be fast but infrequent. Such conformational transitions can be accurately modeled as a stationary point process. In our terminology, the elementary process is then the conformational transition itself. In the application of the point process formalism considered here, the elementary process is water exchange from a buried site, rate-limited by conformational fluctuations. In this case, the point process description is applicable if water exchange is fast and infrequent, regardless of whether the rate-limiting conformational fluctuations can be modeled in terms of discrete states. For water exchange from a well-defined hydration site, the state space is binary because the site is either occupied or not occupied by a given water molecule.

The outline of this paper is as follows. In section 2, we introduce concepts and notation with reference to the continuous-time limit. We then develop algorithms for computing the RCF and SCF, as well as the mean residence and survival times, from the discrete-time data generated by MD simulations. In the Supporting Information, we present fully vectorized Matlab code for these algorithms. Related correlation functions, some of which have been used previously, are discussed. We analyze the systematic binning error associated with the finite sampling resolution and the random statistical error associated with the finite length of the MD trajectory. The details of the error analysis are relegated to the Supporting Information. In the last part of section 2, we discuss the behavior of the SCF in the presence of dynamical disorder. In section 3, we illustrate the theory and algorithms with numerical results for water exchange from an internal site in BPTI. A full account of the exchange kinetics for all four internal hydration sites in BPTI, along with a comparison with experimental data, can be found elsewhere.[24]

2. THEORY

This work was motivated by the need to compare MD simulation data with experimental data on internal-water exchange in proteins, but the following theoretical results are of considerable generality. Throughout section 2, we therefore describe the basic theory in general terms, using the water-exchange case merely as an illustration. However, in sections 2.2 and 2.4, we address several issues that apply specifically to the water-exchange case.

2.1. Residence and Survival Statistics. A *point process* is a particular kind of stochastic process that can be defined with mathematical rigor and considerable generality in measure-theoretic terms.[25] For our purposes, it suffices to define a point process as a random set of points or *events* on the half-line $t > 0$, representing continuous time.[21−23] A realization of such a point process is shown in Figure 2. The essential property that allows a physical process to be modeled as a point process is that the duration of the event is negligibly short compared to a typical interevent interval. The event may be a conformational transition, or, as in our application, an exchange of the water molecule that occupies a cavity within a protein.

We assume that the probability that one event occurs in an infinitesimal time interval δt is proportional to δt and that the probability that more than one event occurs in δt goes to zero faster than δt. Such a point process, where at most one event can occur in a sufficiently small time interval, is sometimes called a *regular* point process. Furthermore, we assume that the point process is *stationary*, meaning that its statistical properties are invariant under translation of the time origin. The stationarity assumption restricts applications to MD simulations of equilibrium systems. In the following, these properties will be implicitly assumed whenever we use the term point process.

A point process may be described in several equivalent ways, depending on how the random variable is defined.[21−23] One description makes use of the random time points at which the events occur; another uses a counting measure defined as the number of events that have occurred in a given time period. For our purposes, it is most convenient to use as the random

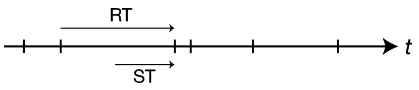


Figure 2. Realization of a point process. Events are represented by tick marks. One residence time (RT) and one survival time (ST) are indicated.

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

variable the *residence time* (RT), that is, the time interval between two successive events (Figure 2). In our application, the RT is the interval between the arrival times of two successive water molecules in the site.

The RT is not the experimentally relevant variable, because the measurement is not synchronized with the events. The random variable that enters in the description of experimental data is the *survival time* (ST), defined as the time interval from an arbitrary initial time (not necessarily coinciding with an event) until the next event (Figure 2). Since the point process formalism has been applied in diverse fields, a rich and potentially confusing terminology has evolved. For example, the ST is variously referred to as the residual lifetime, the persistence time, or the forward recurrence time.

The point process is fully characterized by the RT *probability density* $\psi_R(\tau)$ or, equivalently, by the ST probability density $\psi_S(\tau)$. These functions are normalized as (X = R or S)

$$\int_0^\infty d\tau\ \psi_X(\tau) = 1 \tag{1}$$

For example, $\psi_R(\tau)\ d\tau$ is the probability that the RT is within $d\tau$ of τ. The RT and ST probability densities are in general different, but they are not independent. If we know one of them, we can compute the other (see below).

It is often more convenient to characterize the point process, not by the probability density $\psi_X(\tau)$, but by the integral (X = R or S)

$$Q_X(\tau) \equiv \int_\tau^\infty d\tau' \psi_X(\tau') \tag{2}$$

Since the probability density $\psi_X(\tau)$ must be non-negative, it follows that $Q_X(\tau)$ decreases monotonically with τ from $Q_X(0) = 1$ to $Q_X(\infty) = 0$. Clearly, $Q_X(\tau)$ is the probability that the RT (for X = R) or ST (for X = S) is longer than τ. In probability theory,[26,27] the quantity $1 - Q_X(\tau)$ is known as the (cumulative) distribution function and $Q_X(\tau)$ as the complementary (cumulative) distribution function. Because $Q_X(\tau)$ has the properties of a time correlation function, we refer to $Q_R(\tau)$ as the *residence correlation function* (RCF) and to $Q_S(\tau)$ as the *survival correlation function* (SCF). Again, there is no consensus on terminology in the literature; for example, the RCF $Q_R(\tau)$ is sometimes called the survival (or survivor) function. In our application, $Q_R(\tau)$ is the probability that the duration of a visit to the site by a water molecule (from arrival to departure) exceeds τ, whereas $Q_S(\tau)$ is the probability that a water molecule residing in the site at a given time does not leave the site in the subsequent time interval τ.

Using eq 2 and integrating by parts, we can express the *mean residence time* τ_R and the *mean survival time* τ_S in either of two ways:

$$\tau_X = \int_0^\infty d\tau\ \tau\psi_X(\tau) = \int_0^\infty d\tau\ Q_X(\tau) \tag{3}$$

Since the ST interval generally starts some time after the arrival (Figure 2) one might expect that $\tau_S < \tau_R$. However, for most physically motivated RT densities, the opposite is true: $\tau_S \geq \tau_R$. This is the famous waiting time paradox,[26] which is resolved by noting that τ_R and τ_S are properties of the ensemble of site visits and that the randomly chosen starting point for a ST interval is more likely to fall in a long RT interval than in a short one (length-biased sampling). The significance of τ_R can be appreciated from the fact that $1/\tau_R$ is the average event rate. On average, an MD trajectory of length T contains T/τ_R events.

For our purposes, the key result from probability theory is the link between the RT and ST probability densities (and correlation functions). To establish this link, we consider the probability $\psi_S(\tau)\ d\tau$ that the ST is within $d\tau$ of τ. For a stationary point process, this probability cannot depend on when the ST interval starts. We can therefore choose the starting time at random. Let $\psi_{SR}(\tau|\tau')\ d\tau$ be the conditional probability that the ST is within $d\tau$ of τ, given that the starting time falls within an RT interval of length τ'. The starting time is equally likely to fall anywhere within a given RT, so $\psi_{SR}(\tau|\tau')\ d\tau = \theta(\tau' > \tau)\ d\tau/\tau'$, where $\theta(z)$ is the unit step function (clearly, the ST cannot exceed the RT in which it starts). This conditional probability must now be multiplied by the *a priori* probability that the starting time falls within a RT interval of length τ' (to within $d\tau'$). This is simply the fraction of the trajectory contributed by such RTs, or $\tau'\ \psi_R(\tau')\ d\tau'/\tau_R$. Summing up the contributions from all possible RTs, we thus obtain

$$\psi_S(\tau) = \int_0^\infty d\tau' \frac{\tau'\psi_R(\tau')}{\tau_R}\psi_{SR}(\tau|\tau') = \frac{1}{\tau_R}\int_\tau^\infty d\tau'\psi_R(\tau') = \frac{Q_R(\tau)}{\tau_R} \tag{4}$$

where eq 2 was used in the last step. This heuristic derivation[21] can be made mathematically rigorous.[23] Importantly, the result, eq 4, holds for any stationary point process. Stationarity implies that the individual RTs along the trajectory are identically distributed, but we do not need to assume that they are mutually uncorrelated (as in a renewal process) or exponentially distributed (as in a Poisson process).

The virtue of eq 4 is that it allows us to compute the experimentally relevant SCF $Q_S(\tau)$ from the RT probability density $\psi_R(t)$ deduced from an MD trajectory. Combining eqs 2 and 4, we obtain

$$Q_S(\tau) = \frac{1}{\tau_R}\int_\tau^\infty d\tau' Q_R(\tau') = \frac{1}{\tau_R}\int_\tau^\infty d\tau' \int_{\tau'}^\infty d\tau''\psi_R(\tau'') \tag{5}$$

Several avenues lead to the mean ST τ_S. Using eqs 2−5, we can express τ_S as a double or triple integral involving $\psi_R(\tau)$. However, the most efficient way of computing τ_S makes use of the mean-square residence time, defined as

$$\langle\tau^2\rangle_R \equiv \int_0^\infty d\tau\ \tau^2\ \psi_R(\tau) \tag{6}$$

Using the identity $\psi_R(\tau) = -\tau_R\ d\psi_S(\tau)/d\tau$, which follows by differentiating eq 4, and integrating by parts in eq 6, we obtain the useful identity[23]

$$\langle\tau^2\rangle_R = 2\tau_R\tau_S \tag{7}$$

2.2. Residence Time Histogram. To calculate the SCF $Q_S(\tau)$ by means of eq 5, we need the RT probability density $\psi_R(\tau)$. Accordingly, we first describe how $\psi_R(\tau)$ can be obtained from the MD simulation data in the water-exchange case. For simplicity, we assume that the hydration site is occupied by at most one water molecule. The generalization to a multiply occupied site is outlined in the Supporting Information.

The simulated system contains N_W water molecules per protein molecule, labeled by the index $w = 1, 2, ..., N_W$. The

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

data comprise a series of frames from the MD trajectory, saved at equispaced time points $t_k = k\,\Delta\tau$, with $k = 1, 2, \ldots$ and sampling resolution $\Delta\tau$. The first step is to identify, typically by applying a set of geometric criteria, which water molecule occupies a given hydration site in each frame. We thus obtain an *occupancy vector* $\boldsymbol{A}_0$ where the element $A_0(k)$ is the w index of the current water molecule. If no water molecule satisfies the occupancy criteria, we set $A_0(k) = 0$. We then form a *contracted occupancy vector* $\boldsymbol{A}$ by deleting all zero elements from $\boldsymbol{A}_0$. This deletion has no effect on the RT analysis, which only involves time intervals. If desired, the vacant frames can be subjected to a separate statistical analysis.

Next, we identify the first frame k_0 of each RT in $\boldsymbol{A}$. The difference of adjacent first-frame indices k_0 then gives us a chronological list of the lengths of all RTs, specified as the number n of continuously occupied frames. From this list, we delete the first and last entries, which correspond to incomplete RTs. (However, if $\boldsymbol{A}_0$ begins or ends with a vacant frame, the first or last entry in the RT list are kept since the pruning has then already been accomplished in forming $\boldsymbol{A}$.) The resulting *residence time vector* $\boldsymbol{V}_R = \{n_\alpha\}_{\alpha=1,\ldots,N_R}$ contains all available information about the point process. Ultimately, all statistical properties are derived from $\boldsymbol{V}_R$.

Finally, we bin the RTs into the *residence time histogram* $\boldsymbol{F}_R$, where the element $F_R(n)$ is the number of RTs of length n frames. The total number of RTs is

$$N_R = \sum_{n=1}^{\infty} F_R(n) \tag{8}$$

and the number of unique RTs is equal to the number of nonzero elements in $\boldsymbol{F}_R$. Together, these N_R RTs comprise N_F frames

$$N_F = \sum_{n=1}^{\infty} n\,F_R(n) = \sum_{\alpha=1}^{N_R} n_\alpha \tag{9}$$

In actual computations, all sums over n are truncated after n_{max}, the longest RT contained in $\boldsymbol{F}_R$. Typically, $n_{max} \ll N_F$. Vectorized Matlab code for computing the RT histogram $\boldsymbol{F}_R$ from the occupancy vector $\boldsymbol{A}_0$ is presented in the Supporting Information.

The RT histogram is subject to a systematic binning error related to the finite sampling resolution $\Delta\tau$. As discussed in section 2.5, the binning error is negligible if $N_R \ll N_F$. The mapping from the continuous RT probability density $\psi_R(\tau)$ to the RT histogram then takes the form

$$F_R(n) = N_R\,\psi_R(n\Delta\tau)\,\Delta\tau \tag{10}$$

The RT histogram and the statistics derived from it also have statistical uncertainties related to the finite length of the analyzed trajectory. As discussed in section 2.5, the statistical error is negligible if $N_R \gg 1$. Throughout section 2, except for section 2.5, we ignore both binning error and statistical error.

2.3. Computation of Correlation Functions. The discrete RCF $Q_R(n)$ is obtained by combining the discrete version of eq 2 with eq 10

$$Q_R(n) = \frac{1}{N_R} \sum_{p=n+1}^{\infty} F_R(p) \tag{11}$$

With eq 8, this expression yields the expected result $Q_R(0) = 1$, which also follows, in the continuum limit, from eqs 1 and 2. The mean RT is obtained by combining the discrete version of the second member of eq 3 with eq 10 and then using eq 9

$$\tau_R = \Delta\tau \frac{N_F}{N_R} \tag{12}$$

The discrete SCF $Q_S(n)$ is obtained by combining the discrete version of eq 5 with eq 10 and then using eq 12

$$Q_S(n) = \frac{1}{N_F} \sum_{p=n}^{\infty} \sum_{q=p+1}^{\infty} F_R(q) \tag{13}$$

By rearranging the double sum, this result can be expressed on the simpler, but equivalent, form

$$Q_S(n) = \frac{1}{N_F} \sum_{p=n+1}^{\infty} (p - n)\,F_R(p) \tag{14}$$

With eq 9, this expression yields the expected result $Q_S(0) = 1$, which also follows, in the continuum limit, from eqs 1 and 2. The simplest way of obtaining the mean ST is to combine the discrete version of eq 7 (rather than eq 3) with eq 10 and then using eq 12

$$\tau_S = \frac{\Delta\tau}{2N_F} \sum_{n=1}^{\infty} n^2 F_R(n) \tag{15}$$

Vectorized Matlab code for computing these quantities from the RT histogram $\boldsymbol{F}_R$ is presented in the Supporting Information. Using the Matlab cumsum function, the double sum in eq 13 actually computes faster than the single sum in eq 14, which involves a multiplication.

2.4. Related Algorithms and Correlation Functions. The interpretation of the discrete RCF $Q_R(n)$ as the fraction of RTs that are longer than $n\,\Delta\tau$ suggests that it can be formally expressed as

$$Q_R(n) = \frac{1}{N_R} \sum_{\alpha=1}^{N_R} h_\alpha(n) \tag{16}$$

where the indicator function $h_\alpha(n)$ equals 1 if $n_\alpha > n$ and 0 otherwise. The equivalence of eqs 11 and 16 becomes evident when one recognizes that the sums in these equations are simply two ways of expressing the number of RTs with more than n frames.

In a similar manner, the discrete SCF $Q_S(n)$ can be formally expressed as

$$Q_S(n) = \frac{1}{N_F} \sum_{k=1}^{N_F} g_k(n) \tag{17}$$

where the indicator function $g_k(n)$ equals 1 if $m_k > n$ and 0 otherwise. Here, m_k is the number of frames remaining in the current RT (to which frame k belongs) starting with, and including, frame k. In other words, $g_k(n) = 1$ if the $n + 1$ frames $k, k + 1, \ldots, k + n$ are occupied by the same water molecule. The equivalence of eqs 14 and 17 follows by noting that the sums in these equations are two ways of expressing the total number of frames in all RTs with $n_\alpha > n$ after the first n frames have been removed from each RT.

In the context of water exchange, the indicator function (IF) definition 17 of the SCF was first used by Impey, Madden, and McDonald in their study of the hydration dynamics of simple ions,[20] albeit with an additional averaging over all water molecules in the system to allow for multiple occupancy of the

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

hydration shell. Subsequently, the IF definition has been used in virtually all MD studies of water exchange from the hydration layer of proteins.[10−19]

While simple in concept, the IF algorithm is computationally inefficient. Like the conventional algorithm for computing time correlation functions of continuous configurational variables,[28,29] the IF algorithm involves averaging over time origins. Specifically, a direct evaluation of eq 17 requires us, for a given n, to scan all $N_F - n$ sequences of $n + 1$ consecutive frames in the occupancy vector $\boldsymbol{A}$. For a point process, this is a highly inefficient approach, since all the available kinetic information is contained in the RT histogram $\boldsymbol{F}_R$. As compared to the RT histogram algorithm in eq 13, the IF algorithm in eq 17 involves vastly more terms, since typically $N_F = \text{length}(\boldsymbol{A}) \gg n_{max} = \text{length}(\boldsymbol{F}_R)$. Moreover, eq 17 cannot be fully vectorized. For the water-exchange case studied here, the computation of Q_S from the occupancy vector $\boldsymbol{A}$ is 5 orders of magnitude slower with the IF algorithm than with the RT histogram algorithm (and the Matlab code in the Supporting Information).

Nearly all MD studies of water exchange from the external protein hydration layer have used the SCF $Q_S(\tau)$,[10−19] but a few have used a different correlation function,[30−33] which we refer to as the *total survival correlation function* (tSCF) $Q_{tS}(\tau)$. The tSCF is the probability that the same water molecule resides in the site at the two time points t and $t + \tau$ regardless of whether it resided in the site in the meantime. It is clear that $Q_{tS}(\tau)$ decays more slowly than $Q_S(\tau)$ and that the corresponding mean tST τ_{tS} is longer than τ_S. The discrete tSCF $Q_{tS}(n)$ can be computed by an IF algorithm as in eq 17, but with an indicator function $g^t_k(n)$ that equals 1 if the same water molecule resides in the site in frames k and $k + n$ (regardless of its whereabouts in the meantime) and equals 0 otherwise.[24] It is clear that $Q_{tS}(0) = 1$ and that $Q_{tS}(\infty) = 1/N_W$, where N_W is the number of water molecules in the system. The mean tST is therefore computed as

$$\tau_{tS} = \frac{\Delta\tau}{(N_W - 1)} \sum_{n=0}^{\infty} [N_W Q_{tS}(n) - 1] \tag{18}$$

Close analogs of these correlation functions have been widely used in studies of hydrogen-bond dynamics in liquid water.[34−36] In that context, $Q_S(\tau)$ and $Q_{tS}(\tau)$ are usually referred to as continuous and intermittent correlation functions, respectively.

Another useful quantity is the *total residence correlation function* (tRCF) $Q_{tR}(\tau)$, defined in the same way as $Q_{tS}(\tau)$ except that the time origin now coincides with an exchange event. The discrete tRCF $Q_{tR}(n)$ can be computed as

$$Q_{tR}(n) = \frac{1}{N_R} \sum_{k_0} g^t_{k_0}(n) \tag{19}$$

where the IF $g^t_{k_0}(n)$ equals 1 if the same water molecule resides in the site in frames k_0 and $k_0 + n$ (regardless of its whereabouts in the meantime) and equals 0 otherwise. The sum now runs only over the indices k_0 of the first frame in each of the N_R RTs in the trajectory. As for the tSCF, we have $Q_{tR}(0) = 1$ and $Q_{tR}(\infty) = 1/N_W$, and the mean tRT τ_{tR} is computed from eq 18 with $Q_{tS}(n)$ replaced by $Q_{tR}(n)$.

For water exchange from the external hydration layer of a protein, the SCF $Q_S(\tau)$ and the mean ST τ_S are sensitive to the demarcation of the hydration site. Since τ_S is in the picosecond range for the vast majority of surface sites,[37] the sampling interval $\Delta\tau$ must be of order 100 fs or shorter to keep the binning error small ($N_F \gg N_R$). At this resolution, subpicosecond intermolecular vibrations are observed as frequent recrossings of the site boundary, leading to a shorter τ_S. The recrossing problem can be minimized by ignoring excursions from the site that are shorter than some cutoff value $\delta\tau$.[20] Alternatively, one may compute the more robust mean tST τ_{tS}, which is less sensitive to the site demarcation details and does not depend on the somewhat arbitrary parameter $\delta\tau$.[33]

Exchange of internal water molecules involves large free energy barriers, so τ_S is much longer (typically 10 ns−10 μs) than for the external hydration layer. For this reason, internal-water exchange can be studied directly by MRD experiments.[7] The MRD experiment probes the tSCF $Q_{tS}(\tau)$ but, because of the large barrier, an exchanged water molecule is not likely to revisit the site significantly more often than any other water molecule. Since exchange removes essentially all correlation with the site, we expect that $Q_{tS}(\tau) = Q_S(\tau)$. The accuracy of this approximation can be quantitatively assessed with the aid of MD data.[24]

2.5. Error Analysis. The RT and ST statistics considered above are subject to two kinds of error that we refer to as *binning error* and *statistical error*. A detailed analysis of these errors can be found in the Supporting Information; here we merely summarize the main results.

The systematic binning error is caused by two effects, both related to the finite sampling resolution $\Delta\tau$. The *discretization error* is the result of replacing the continuous RT probability density $\psi_R(\tau)$ by a discrete RT histogram $F_R(n)$. This error is of second order in $\Delta\tau/\tau_R$

$$F_R(n) = N^0_R\, \Delta\tau\, \psi_R(n\Delta\tau) \left[1 + O\left(\frac{\Delta\tau}{\tau_R}\right)^2\right] \tag{20}$$

where N^0_R is the number of RTs that would have been observed in the trajectory in the limit $\Delta\tau \to 0$. In general, N^0_R exceeds the number N_R of RTs observed with a finite resolution $\Delta\tau$, because RTs shorter than $\Delta\tau$ may escape detection. This *resolution error* is of first order in $\Delta\tau/\tau_R$:

$$N_R = N^0_R \left[1 - O\left(\frac{\Delta\tau}{\tau_R}\right)\right] \tag{21}$$

The ST-related quantities $Q_S(n)$ and τ_S in eqs 13 and 15 do not involve N_R and are therefore only affected by the discretization error. These quantities are thus accurate to first order in $\Delta\tau/\tau_R$. But the RT-related quantities $Q_R(n)$ and τ_R in eqs 11 and 12 do involve N_R and are therefore also affected by the resolution error. These quantities are therefore accurate only to zeroth order in $\Delta\tau/\tau_R$. The higher accuracy of the ST-related quantities can be understood by noting that they are biased toward the longer RTs and therefore are insensitive to the loss of short RTs.

The random statistical error results from the finite length of the trajectory, leading to an incomplete sampling of the RT ensemble. Our primary data are the $\boldsymbol{V}_R$ vector, a sequence of N_R RT intervals. The number N_R is thus a measure of the amount of information at our disposal. If we sample the trajectory more densely by decreasing $\Delta\tau$, we increase the number N_F of frames, but we do not gain more information. The statistical error should therefore depend on N_R but not on N_F.

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Specifically, if the RTs are mutually uncorrelated, we expect[38] the statistical error to be proportional to $N_R^{-1/2}$.

The mean RT τ_R obtained from eq 12 is an unbiased estimator of the ensemble average $\langle\tau_R\rangle$ for an ensemble of V_R vectors, each with N_R RTs. The statistical error in τ_R, which is an unbiased estimator of the standard deviation, can be expressed as

$$s(\tau_R) = \left[\frac{\tau_R(2\tau_S - \tau_R)}{N_R - 1}\right]^{1/2} \tag{22}$$

This result is strictly valid only if the RTs are mutually uncorrelated. If the RTs are correlated, eq 22 only yields a lower bound on the statistical error. A more accurate estimate may then be obtained by the block renormalization method.[39]

The statistical error in the RCF $Q_R(n)$ is, in the absence of RT correlations, given by

$$s[Q_R(n)] = \left[\frac{Q_R(n)[1 - Q_R(n)]}{N_R - 1}\right]^{1/2} \tag{23}$$

This result differs from the standard formula for the statistical error in a time correlation function of a Gaussian random variable,[40] which in our notation is roughly $(2/N_R)^{1/2}[1 - Q_R(n)]$.

For the ST-related quantities τ_S and $Q_S(n)$, the error analysis is more involved. In the Supporting Information, we present approximate results for the statistical error in these quantities in the absence of RT correlations. To our knowledge, no error analysis has been presented previously for any of the four quantities considered here, with the exception of a proposal[11] that $s[Q_S(n)] = \{Q_S(n)\ [1 - Q_S(n)]/N_F\}^{1/2}$. We believe that this result is incorrect, because, as argued above, the error should depend on N_R rather than on N_F.

2.6. Dynamical Disorder. In general, several inactive (I) and active (A) conformational states are thermally accessible to the protein. In the conformational gating regime, the observed elementary process (e.g., internal-water exchange) only provides information about the I states. Let there be N_Γ I states Γ_i, each characterized by a relative population f_i and an I → A transition rate k_i. Without loss of generality, we can assume that this rate is constant in time. (The states can always be defined so that this condition is satisfied.) This implies (see below) that $\tau_{S,i} = \tau_{R,i} = 1/k_i$.

Different physical scenarios can be envisaged where multiple I states must be taken into account. The protein might fluctuate among different conformational substates with different k_i rates, or water molecules in a multiply occupied cavity might interchange among subsites with different k_i rates. Such dynamical disorder can introduce correlations among the RTs in the trajectory. For example, there might be a cluster of short RTs in one part of the trajectory and a cluster of long RTs in another part.

Dynamical disorder can be formally described by a stochastic rate equation with an effective rate constant $k(\Gamma)$ that depends on the fluctuating conformational state Γ

$$\frac{d}{d\tau}Q_S(\tau) = -k(\Gamma)Q_S(\tau) \tag{24}$$

Rather than discussing the general case,[5,27] we shall focus on three important special cases. The simplest case is when the rate constant k does not fluctuate at all. The stationary point process then reduces to a Poisson process, and it follows from eqs 24 and 1−3 that

$$Q_S(\tau) = Q_R(\tau) = \frac{\psi_S(\tau)}{k} = \frac{\psi_R(\tau)}{k} = \exp(-k\tau) \tag{25}$$

and that

$$\tau_S = \tau_R = \frac{1}{k} \tag{26}$$

Another special case is the fast fluctuation limit, where the transitions among Γ states are much faster than the difference of the corresponding k_i rates. The solution to eq 24 is then again an exponential

$$Q_S(\tau) = \exp[-\langle k(\Gamma)\rangle\tau] \tag{27}$$

so that, with eq 3

$$\tau_S = \frac{1}{\langle k(\Gamma)\rangle} = \left[\sum_{i=1}^{N_\Gamma}\frac{f_i}{\tau_{S,i}}\right]^{-1} \tag{28}$$

As can be seen from eqs 1−3, the exponential form of $Q_S(\tau)$ in eq 27 implies that the point process reduces to a Poisson process also in this case.

Of greater interest is the slow fluctuation ('static' disorder) limit, where the transitions among Γ states are much slower than the difference of the corresponding k_i rates. The solution to eq 24 is then a linear combination of exponentials

$$Q_S(\tau) = \sum_{i=1}^{N_\Gamma} f_i \exp(-\tau/\tau_{S,i}) \tag{29}$$

so that, with eq 3

$$\tau_S = \sum_{i=1}^{N_\Gamma} f_i\tau_{S,i} \tag{30}$$

In this limit, τ_S thus tends to be dominated by the slowest state (with the longest $\tau_{S,i}$), whereas the opposite is true in the fast fluctuation limit.

Even if the experimentally observed SCF $Q_S(\tau)$ appears to be exponential, one cannot rule out static disorder. To see this, consider the bi-Poissonian RCF

$$Q_R(\tau) = f_1 \exp(-\tau/\tau_{R,1}) + (1 - f_1) \exp(-\tau/\tau_{R,2}) \tag{31}$$

which, according to eqs 2−4, corresponds to the SCF

$$Q_S(\tau) = c_1 \exp(-\tau/\tau_{R,1}) + (1 - c_1) \exp(-\tau/\tau_{R,2}) \tag{32}$$

with

$$c_1 = \frac{f_1\tau_{R,1}}{f_1\tau_{R,1} + (1 - f_1)\tau_{R,2}} \tag{33}$$

Now assume that $f_1\tau_{R,1} \ll (1-f_1)\tau_{R,2}$ so that $c_1 \ll 1$. Then $Q_S(\tau)$ will appear exponential at all times with $\tau_S \approx \tau_{R,2}$, whereas (the experimentally inaccessible) RCF $Q_R(\tau)$ may exhibit a substantial (if f_1 is not too small) initial decay on the short time scale $\tau_{R,1}$ and a tail that decays on the same long time scale $\tau_{R,2}$ as $Q_S(\tau)$.

3. NUMERICAL RESULTS

3.1. Simulation Data. The 1.031 ms trajectory analyzed here, kindly provided by D. E. Shaw Research, pertains to a classical MD simulation at 300 K of the protein bovine

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

pancreatic trypsin inhibitor (BPTI) solvated by 4215 water molecules and 6 chloride ions.[9] A variant of the AMBER ff99SB force field was used, along with the TIP4P-Ew water model. Further details can be found in the original publication.[9] The analyzed trajectory comprises 4,125,000 frames with sampling resolution $\Delta\tau$ = 0.25 ns.

According to crystallography,[41] the BPTI molecule contains four singly occupied internal hydration sites, traditionally denoted W111, W112, W113, and W122. The first three sites are located in a pore formed by two extended polypeptide loops. The water molecules occupying these sites are mutually hydrogen-bonded, with W113 at the bottom of the pore. A full analysis of the water exchange kinetics for all four sites, including a comparison with experimental results,[7] is presented elsewhere.[24] Here, as an illustrative application of the theoretical results presented in section 2, we only consider the deeply buried site W113. As a contrast, we also briefly consider the external hydration site W143, located in a moderately deep surface pocket. The geometric criteria used to construct the occupancy vector $\boldsymbol{A}_0$ are fully described elsewhere.[24] Briefly, the internal site W113 is taken to be occupied if the water molecule makes at least 2 H-bonds to the 3 site-defining protein atoms Y10.O, K41.N, and N44.N, with the H-bond cutoffs $R(\mathrm{H}\cdots\mathrm{A}) \leq 3.0$ Å and $\theta(\mathrm{DHA}) \geq 130°$.

According to the simulation, hydration site W113 in BPTI is vacant 13.3% of the time. In contrast, crystallography[41] and NMR[7] both indicate full occupancy of this site. This discrepancy can be attributed to a minor (a few k_BT) force field deficiency, distorting the conformational distribution. For example, previous analyses of this trajectory[9,42] found substantial differences from experiment[43,44] in the populations of the conformational states associated with rotational isomers of the C14 − C38 disulfide bond. This discrepancy need not concern us here, since our objective is to illustrate the theory and algorithms rather than to compare with experiment.

The force field error can be regarded as a pedagogical asset here, since it allows us to exhibit more clearly the effects of dynamical disorder. In the following, we present results for the entire trajectory, including all conformational states, as well as for a particular conformational state, denoted M1, which is populated in 25% of the trajectory. According to experiment, state M1 corresponds to the dominant (∼95%)[44] C14−C38 isomer. In the simulated M1 state, the occupancy of site W113 is 99.92%.

Because state transitions do not in general coincide with water exchange events, a convention must be adopted for extracting a state-specific subtrajectory for a given site. The simplest procedure is to concatenate all frames belonging to the selected state. However, the resulting truncation of RTs introduces a bias that shifts the RT distribution to shorter values. We therefore use a 'democratic' approach, where, for a given site, each RT is assigned to the state that is represented in the largest number of frames in that RT. A detailed analysis of water exchange events from site W113 that occur when the protein is in disulfide state M1 shows that the protein remains in state M1 during and after the exchange in 90% of the exchange events, whereas a brief visit (a transition state with lifetime <5 ns) to state M2 is seen in 10% of the exchange events.[24] In either case, the exchange event thus links two RTs 'democratically' classified as belonging to state M1.

3.2. Global Water Exchange Kinetics. We begin by analyzing all the 3.56 million frames where site W113 is occupied. These frames can be subdivided into groups corresponding to different conformational states of the BPTI molecule. However, in this subsection, we analyze all frames without regard to conformational state. Relevant statistics of the $\boldsymbol{A}$, $\boldsymbol{V}_R$, and $\boldsymbol{F}_R$ vectors are collected in Table 1.

Table 1. Statistics of $\boldsymbol{A}$, $\boldsymbol{V}_R$, and $\boldsymbol{F}_R$ Vectors

site state	W113 all	W113 M1	W113 M1[a]	W143 M1
N_F	3 560 059	1 061 268	1 059 599	294 816
unique waters	4213	132	53	4215
N_R	37 499	144	60	84 225
unique RTs	675	93	60	48
τ_R[b]	24 ± 2 ns	1.8 ± 0.3 μs	4.4 ± 0.5 μs	0.88 ± 0.25 ns
τ_S[b]	2.5 ± 0.2 μs	4.1 ± 0.5 μs	4.1 ± 0.5 μs	0.93 ± 0.08 ns

[a]RTs shorter than 50 ns (n < 200) omitted. [b]The quoted uncertainty is the uncorrelated statistical error for W113 and an order-of-magnitude estimate of the binning error for W143.

During the course of the 1 ms simulation, internal site W113 is visited by 4213 unique water molecules, all but two of the 4215 water molecules in the system. These visitors make N_R = 37,499 visits to the site. Most visits are brief: 98% are less than 100 frames, and 43% of the visits last only one frame. Therefore, the number of unique RTs (675) is much smaller than N_R. In addition, there is an unknown number of visits that are too short to be observed at our resolution, $\Delta\tau$ = 0.25 ns. The observed RTs span 5 orders of magnitude: from 0.25 ns to 19 μs. The longest RT determines the length of the frequency vector, $n_{max} = \tau_{max}/\Delta\tau$ = 76,749, but only 675 of these elements (<1%) are nonzero.

Figure 3 displays the RT vector $\boldsymbol{V}_R$ on a semilog format and, to the right of it, a logarithmically binned version of the RT histogram $\boldsymbol{F}_R$. (The first log bin corresponds to $F_R(1)$, but each of the other log bins includes several n values.) At the top of Figure 3, the RTs for site W113 are assigned to conformational states. (The five named states in Figure 3 all refer to rotational isomers of the C14−C38 disulfide bond.[9,42−44]) The experimentally dominant M1 state is populated in 25.4% of all frames and in 29.0% of all frames where site W113 is occupied, but only 0.4% of the RTs for site W113 are assigned to state M1 (Figure 3). This is so because most of the time spent in state M1 is contributed by long RTs, while most RTs are short.

The residence and survival correlation functions, $Q_R(\tau)$ and $Q_S(\tau)$, for site W113 are shown in Figure 4. The dashed curve was computed by Shaw et al., who referred to it as the 'survival probability'.[9] That curve was computed by the so-called Kaplan−Meier algorithm (P. Maragakis, personal communication), which is widely used in the medical field to correct for patient losses during clinical trials.[45] However, this (inefficient) algorithm does not yield the experimentally relevant SCF but rather the RCF $Q_R(\tau)$. The substantial deviation from our $Q_R(\tau)$ in Figure 4 can be attributed to the different methods used to define water occupancy in the site. Specifically, Shaw et al. used two different hydrogen bond cutoffs (4.0 and 8.0 Å), thereby eliminating the shortest RTs, much as in the modified IF algorithm[20] (see section 2.4).

The RCF and SCF are clearly very different: $Q_R(\tau)$ decays largely on a nanosecond time scale, whereas $Q_S(\tau)$ decays largely on a microsecond time scale (Figure 4). However, $Q_R(n)$ has a weak tail extending beyond 1 μs, and $Q_S(n)$ decays slightly even below 1 ns (Figure 4). The mean RT and ST

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

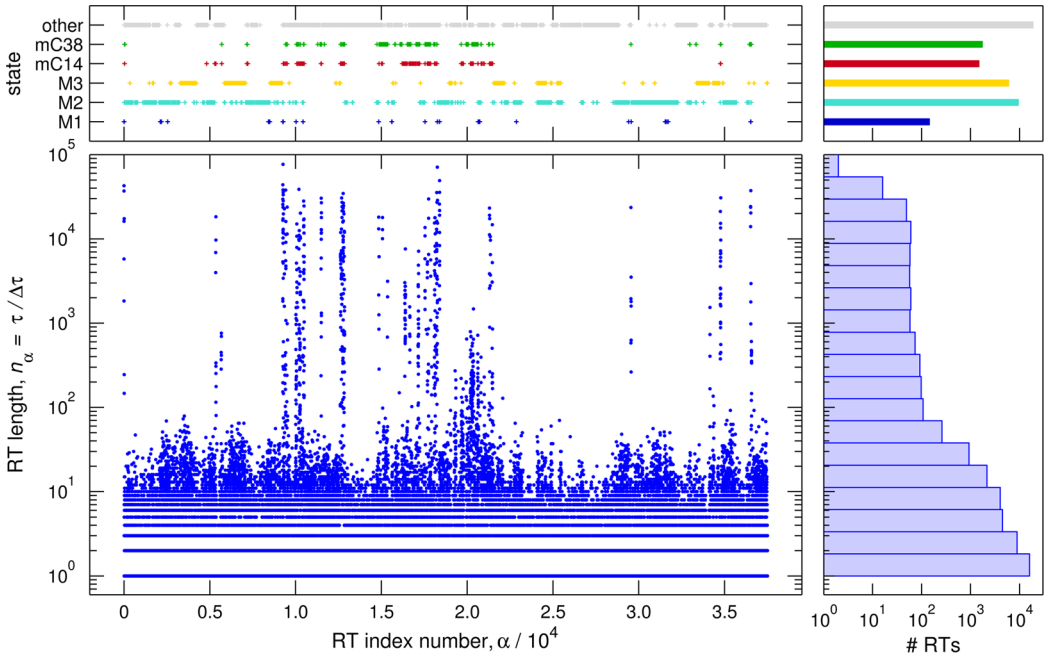


Figure 3. Residence time (RT) vector $\boldsymbol{V}_R$ for site W113 (main panel), with a logarithmically binned RT histogram to the right. The top panels show the assignment of RTs to states along the trajectory (left) and the number of RTs belonging to different states (right).

therefore differ by 'only' 2 orders of magnitude (Table 1). But, the essential observations are that $\tau_S \gg \tau_R$ and that $Q_R(\tau)$ is highly nonexponential. These are the hallmarks of dynamical disorder.

Experimentally, we cannot probe $Q_R(\tau)$ directly so the only way to detect dynamical disorder is via the multiexponential decay of $Q_S(\tau)$. In virtually all cases of interest, $Q_S(\tau)$ can be expressed as a linear combination of exponentials. For example, the general solution of eq 24 has this form. Bearing in mind that it is only in the slow fluctuation limit that the parameters can be interpreted as state populations and state-dependent mean survival times, we can always represent $Q_S(\tau)$ by eq 29. In principle, the parameters in eq 29 can be determined by any nonlinear optimization method. In practice, this can be problematic because of convergence to local minima, dependence on initial parameter estimates, and the need for an *a priori* assumption about the number of exponentials. These problems are avoided in linear inverse methods, such as the non-negative least-squares (NNLS) algorithm.[46,47]

So as not to place undue weight on the most densely sampled long-τ part of $Q_S(\tau)$, we resample by selecting 100 points uniformly spaced along the curve in a semilog plot (Figure 5). For the NNLS kernel (or basis set), we use 200 exponentials with log-spaced decay times between 2.5 ns and 25 μs. NNLS inversion yields 10 components with nonzero amplitude, which is reduced to 6 after merging components that differ by less than 10% in decay time (Figure 6). A reconstruction of $Q_S(\tau)$ using eq 29 with the parameter values of these 6 components faithfully reproduces the data (Figure 5). The reconstructed mean ST obtained from eq 22, τ_S(NNLS) = 2.455 μs, is nearly identical to that computed from eq 15 using the RT histogram, τ_S = 2.459 μs.

The dominant component in the NNLS deconvolution has amplitude f_1 = 0.877 and decay time $\tau_{S,1}$ = 2.78 μs. The second largest component has f_1 = 0.064 and $\tau_{S,1}$ = 0.26 μs. The remaining 4 components all have small amplitudes (<0.025) and short decay times (0.3−32 ns), but it is these fast small-

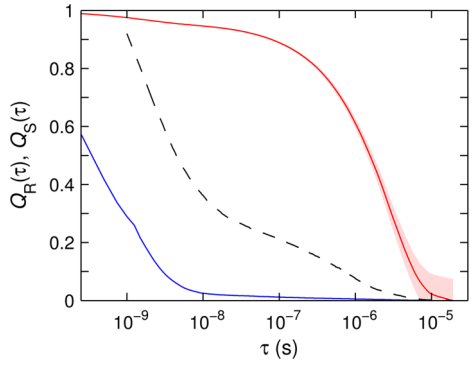


Figure 4. Residence correlation function (RCF) $Q_R(\tau)$ (blue) and survival correlation function (SCF) $Q_S(\tau)$ (red) for site W113 in all states. The shaded band (not visible for the RCF) represents the statistical error. The dashed curve is the RCF computed by Shaw et al.[9]

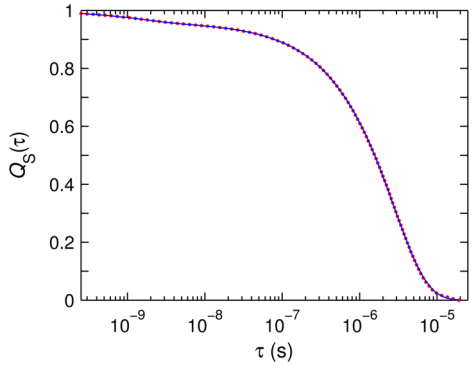


Figure 5. Resampled survival correlation function $Q_S(\tau)$ for site W113 in all states (dots) and NNLS reconstruction (curve).

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

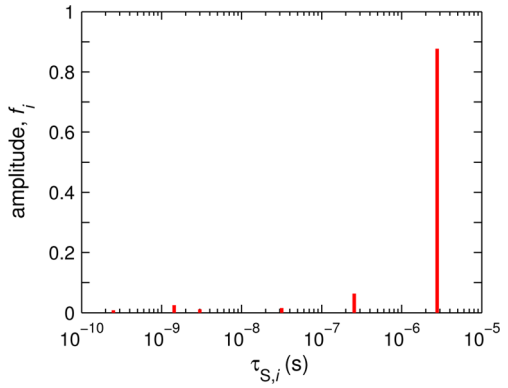


Figure 6. NNLS components in survival correlation function $Q_S(\tau)$ for site W113 in all states. Shown are the amplitude f_i and mean survival time $\tau_{S,i}$ of component i in the decomposition of $Q_S(\tau)$ as in eq 29.

amplitude components that are responsible for the dramatic difference between $Q_S(\tau)$ and $Q_R(\tau)$ in Figure 4 and between τ_S and τ_R (Table 1). From an experimental point of view, these fast components would go undetected, since they have virtually no effect on $Q_S(\tau)$ or τ_S.

How the very weak long-τ tail of $Q_R(\tau)$ can 'give rise' to the dominant microsecond decay of $Q_S(\tau)$ can be appreciated with the aid of eqs 32 and 33. An NNLS deconvolution of $Q_R(\tau)$ yields 4 components, which can be approximated as one fast component with f_1 = 0.98 and $\tau_{R,1}$ = 1.0 ns and one slow component with f_2 = 0.02 and $\tau_{R,2}$ = 1.2 μs. With these numbers, eq 33 yields c_1 = 0.04 for the amplitude of the fast exponential in $Q_S(\tau)$. Therefore, $Q_S(\tau)$ appears to decay almost exponentially with a decay time τ_S close to $\tau_{R,2}$.

3.3. State-Specific Water Exchange Kinetics. We now restrict the analysis of site W113 to the subset of ~1 million frames belonging to RTs where M1 is the most populated conformational state. The RT vector $\boldsymbol{V}_R$ is shown in Figure 7, and relevant statistics are collected in Table 1. Comparing with the all-states RT vector, the number N_R of site visits has dropped by 2 orders of magnitude to merely 144. While the RT range is the same as before, there are far fewer short visits to site W113 when the protein is in the M1 state. Therefore, τ_S now exceeds τ_R by only a factor of ~2 (Table 1). (If M1-state-specific RTs are extracted by simple concatenation rather than 'democratically', τ_R is further shortened by a factor 0.64, whereas τ_S is reduced by merely 1%.) Nevertheless, $Q_R(\tau)$ still differs markedly from $Q_S(\tau)$, although they seem to converge at long τ values (Figure 8).

The logarithmically binned RT histogram for site W113 in state M1 is clearly bimodal, with maxima at 0.25 ns and 5 μs (Figure 7). The subset of short RTs, which is responsible for the remaining difference between $Q_R(\tau)$ and $Q_S(\tau)$ (see below), is much less prominent for the other two deeply buried hydration sites (W112 and W122).[24] For site W113, 19 of the 144 RTs in state M1 last only a single frame, whereas the other two sites have no single-frame RTs.[24] The shortest RTs may reflect transient violations of the H-bond criteria defining occupancy in site W113, perhaps during the actual exchange, but they have virtually no effect on the experimentally relevant ST statistics (Table 1).

Analyzing the 'slow' subset of RTs for site W113 in state M1, including only the 60 RTs with 200 frames (50 ns) or more, we find, as expected, that $Q_R(n)$ and $Q_S(n)$ no longer differ significantly (Figure 9). Within the rather large statistical errors

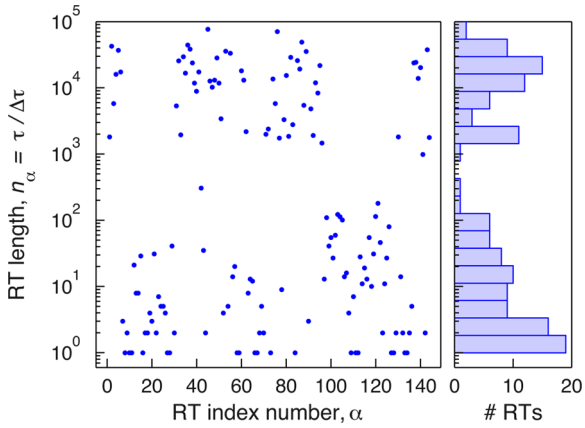


Figure 7. Residence time (RT) vector $\boldsymbol{V}_R$ for site W113 in state M1 (main panel), with a logarithmically binned RT histogram to the right.

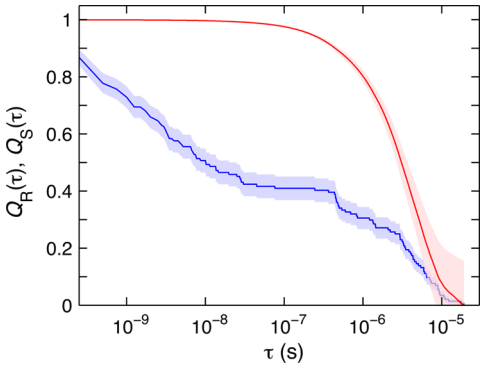


Figure 8. Residence correlation function (RCF) $Q_R(\tau)$ (blue) and survival correlation function (SCF) $Q_S(\tau)$ (red) for site W113 in state M1. The shaded bands represent the statistical error.

(N_R is now only 60), τ_R and τ_S also agree (Table 1). NNLS deconvolution now yields a single component with τ_S(NNLS) = 4.3 μs, consistent with τ_S = 4.1 ± 0.5 μs as obtained directly from the RT histogram. We therefore conclude that I → A conformational gating in state M1 can be accurately described as a Poisson process.

While most of the external protein hydration layer exhibits picosecond water dynamics, less than a factor 2 slower than in bulk water,[37] nanosecond water exchange times are indicated

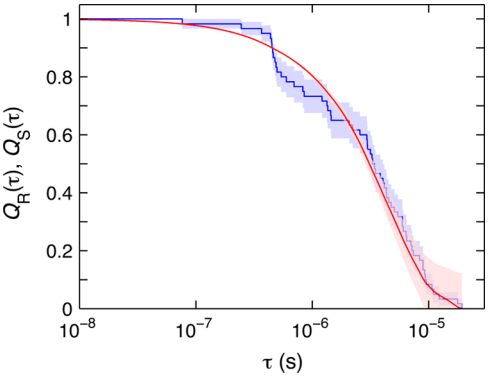


Figure 9. Residence correlation function (RCF) $Q_R(\tau)$ (blue) and survival correlation function (SCF) $Q_S(\tau)$ (red) for site W113 in state M1, including only RTs of at least 50 ns. The shaded bands represent the statistical error.

dx.doi.org/10.1021/ct400161u | J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

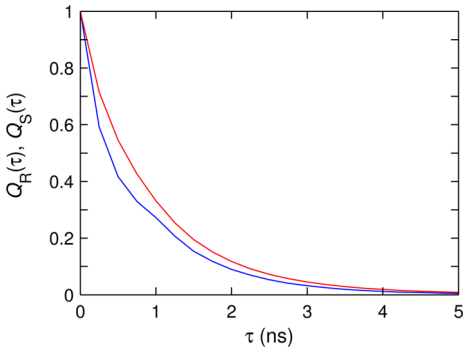


Figure 10. Residence correlation function (RCF) $Q_R(\tau)$ (blue) and survival correlation function (SCF) $Q_S(\tau)$ (red) for external hydration site W143 in state M1. The statistical error is within the line thickness, but the discretization effect is evident in $Q_R(\tau)$.

for a small number of external hydration sites, confined in deep pockets on the protein surface.[12,15,37] An example from this category is site W143 in BPTI. For this site, in the experimentally dominant M1 state, $Q_R(\tau)$ and $Q_S(\tau)$ decay nearly exponentially (Figure 10), and τ_R and τ_S are both ∼1 ns (Table 1). NNLS deconvolution of $Q_S(\tau)$ for site W143 in the M1 state yields a dominant component with $f_1 = 0.87$ and $\tau_{S,1} = 1.0$ ns. To a good approximation, water exchange from a confined external hydration site can thus be described as a Poisson process.

We close this section with two technical remarks. When N_R is not very large, the steps in $Q_R(\tau)$ are clearly visible, whereas $Q_S(\tau)$ still appears smooth (see Figures 8 and 9). This behavior is readily understood from eqs 11 and 13, which yield

$$\Delta Q_R(n) \equiv Q_R(n-1) - Q_R(n) = \frac{F_R(n)}{N_R}$$
$$\Delta Q_S(n) \equiv Q_S(n-1) - Q_S(n) = \frac{N_R}{N_F} Q_R(n-1) \quad (34)$$

The length of the RT histogram vector, determined by the longest RT, is $n_{max} = 76{,}749$ for site W113 in all three cases considered here, but the number of nonzero elements of this vector is in the range 60−675 (Table 1). The RCF $Q_R(n)$ therefore exhibits long (at least on a linear scale) plateau regions, where $F_R(n) = 0$, interrupted by steps of magnitude $1/N_R$ (except for very small n, $F_R(n)$ does not exceed 1). If N_R is not very large, these steps will be apparent. In contrast, $Q_S(n)$ does not show any plateau regions but decreases at every n by a very small amount proportional to $N_R/N_F = \Delta\tau/\tau_R$.

The second technical issue concerns the finding that $\tau_S < \tau_R$ for the 'slow' RT subset for site W113 in state M1 (Table 1). Whereas $\tau_S \geq \tau_R$ is the 'normal' behavior, expected because of the sampling bias (Sect. 2.1), the reverse situation can occur if the shortest RTs are removed from the RT distribution. Consider the normalized RT probability density

$$\psi_R(\tau) = \begin{cases} 0, & \text{for } \tau < \tau_0 \\ \dfrac{1}{\hat{\tau}_R} \exp\left[-\dfrac{(\tau-\tau_0)}{\hat{\tau}_R}\right], & \text{for } \tau \geq \tau_0 \end{cases} \quad (35)$$

For this truncated RT distribution, eqs 3 and 7 yield $\tau_R = \hat{\tau}_R + \tau_0$ and $\tau_S = \hat{\tau}_R + \tau_0^2/[2(\hat{\tau}_R + \tau_0)]$, showing that the ratio τ_S/τ_R goes from 1 for $\tau_0 \ll \hat{\tau}_R$ to 1/2 for $\tau_0 \gg \hat{\tau}_R$.

4. CONCLUSIONS

As advances in molecular simulation technology provide access to longer time scales and larger systems, there is an increasing demand for general theoretical models and efficient numerical algorithms that can be used to make sense of the vast amounts of computer-generated data. Here, we have demonstrated that the stochastic theory of stationary point processes provides a useful framework for efficient analysis of MD simulation data on intermittent dynamics, such as infrequent transitions among discrete conformational states in proteins, as well as elementary processes that are rate-limited (gated) by such dynamics. We have illustrated the point process approach by an application where exchange of a buried water molecule constitutes the elementary process, but the same approach is applicable to a wide range of dynamical processes.

The key element of the point process theory is the rigorous link between the directly computed residence times and the experimentally accessible survival times. This link provides conceptual insight, and it can be used to speed up the computation of survival time statistics by orders of magnitude. Dynamical disorder is the rule rather than the exception for proteins and other complex macromolecular systems. We have therefore discussed how dynamical disorder can be diagnosed and analyzed. We have also presented a detailed analysis of the binning and statistical errors associated with the residence and survival correlation functions and the mean residence and survival times. Previous MD analyses of the survival times and correlation functions for water exchange have not addressed this important issue.

Our aim here has been to present a coherent theoretical framework that can be applied, adapted, and extended for the analysis of diverse dynamical phenomena in proteins and other complex systems. As regards internal-water exchange in proteins, very little MD work has been carried out to date since trajectories of the required length were not available. The computational part of the present work made use of a 1 ms MD trajectory of fully hydrated BPTI.[9] The complete analysis of all four internal water molecules in BPTI, providing insights about rare conformational fluctuations and a valuable benchmark on the accuracy of the force field, will appear elsewhere.[24]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Matlab code for computing the RT vector and histogram and the RT and ST statistics with errors, a brief discussion of multiply occupied hydration sites, and a detailed analysis of the binning and statistical errors in the RT and ST statistics. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: bertil.halle@bpc.lu.se.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank D. E. Shaw Research for sharing the BPTI trajectory, Paul Maragakis for valuable communication, Shuji Kaieda for helpful comments on the manuscript, and the Swedish Research Council for financial support.

REFERENCES

(1) Hvidt, A.; Nielsen, S. O. *Adv. Prot. Chem.* **1966**, *41*, 287−386.
(2) Woodward, C.; Simon, I.; Tüchsen, E. *Mol. Cell. Biochem.* **1982**, *48*, 135−160.
(3) Szabo, A.; Shoup, D.; Northrup, S. H.; McCammon, J. A. *J. Chem. Phys.* **1982**, *77*, 4484−4493.
(4) Agmon, N.; Hopfield, J. J. *J. Chem. Phys.* **1983**, *78*, 6947−6959.
(5) Zwanzig, R. *Acc. Chem. Res.* **1990**, *23*, 148−152.
(6) Zhou, H.-X.; Wlodek, S. T.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9280−9283.
(7) Persson, E.; Halle, B. *J. Am. Chem. Soc.* **2008**, *130*, 1774−1787.
(8) Nilsson, T.; Halle, B. *J. Chem. Phys.* **2012**, *137* (054503), 1−15.
(9) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.
(10) Garcia, A. E.; Stiller, L. *J. Comput. Chem.* **1993**, *14*, 1396−1406.
(11) Rocchi, C.; Bizzarri, A. R.; Cannistraro, S. *Chem. Phys.* **1997**, *214*, 261−276.
(12) Garcia, A. E.; Hummer, G. *Proteins* **2000**, *38*, 261−272.
(13) Luise, A.; Falconi, M.; Desideri, A. *Proteins* **2000**, *39*, 56−67.
(14) Sterpone, F.; Ceccarelli, M.; Marchi, M. *J. Mol. Biol.* **2001**, *311*, 409−419.
(15) Henchman, R. H.; Tai, K.; Shen, T.; McCammon, J. A. *Biophys. J.* **2002**, *82*, 2671−2682.
(16) Massi, F.; Straub, J. E. *J. Comput. Chem.* **2003**, *24*, 143−153.
(17) Dastidar, S. G.; Mukhopadhyay, C. *Phys. Rev. E* **2003**, *68* (021921), 1−9.
(18) Hua, L.; Huang, X.; Zhou, R.; Berne, B. J. *J. Phys. Chem. B* **2006**, *110*, 3704−3711.
(19) Sengupta, N.; Jaud, S.; Tobias, D. J. *Biophys. J.* **2008**, *95*, 5257−5267.
(20) Impey, R. W.; Madden, P. A.; McDonald, I. R. *J. Phys. Chem.* **1983**, *87*, 5071−5083.
(21) Cox, D. R.; Miller, H. D. *The Theory of Stochastic Processes*; Chapman & Hall: London, 1965.
(22) Cox, D. R.; Isham, V. *Point Processes*; Chapman & Hall: London, 1980.
(23) Daley, D. J.; Vere-Jones, D. *An Introduction to the Theory of Point Processes*; Vol. *I*: *Elementary Theory and Methods*, 2nd ed.; Springer: New York, 2003.
(24) Persson, F.; Halle, B. *J. Am. Chem. Soc.*, in press.
(25) Daley, D. J.; Vere-Jones, D. *An Introduction to the Theory of Point Processes*; Vol. *II*: *General Theory and Structure*, 2nd ed.; Springer: New York, 2008.
(26) Feller, W. *An Introduction to Probability Theory and Its Applications*, 2nd ed.; Wiley: New York, 1967; Vol. *II*.
(27) van Kampen, N. G. *Stochastic Processes in Physics and Chemistry*, 3rd ed.; Elsevier: Amsterdam, 2007.
(28) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford Univ. Press: Oxford, 1987.
(29) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: London, 2002.
(30) Brunne, R. M.; Liepinsh, E.; Otting, G.; Wüthrich, K.; van Gunsteren, W. F. *J. Mol. Biol.* **1993**, *231*, 1040−1048.
(31) Abseher, R.; Schreiber, H.; Steinhauser, O. *Proteins* **1996**, *25*, 366−378.
(32) Makarov, V. A.; Andrews, B. K.; Smith, P. E.; Pettitt, B. M. *Biophys. J.* **2000**, *79*, 2966−2974.
(33) Schröder, C.; Rudas, T.; Boresch, S.; Steinhauser, O. *J. Chem. Phys.* **2006**, *124* (234907), 1−18.
(34) Stillinger, F. H. *Adv. Chem. Phys.* **1975**, *31*, 1−101.
(35) Rapaport, D. C. *Mol. Phys.* **1983**, *50*, 1151−1162.
(36) Luzar, A. *J. Chem. Phys.* **2000**, *113*, 10663−10675.
(37) Mattea, C.; Qvist, J.; Halle, B. *Biophys. J.* **2008**, *95*, 2951−2963.
(38) Bevington, P. R.; Robinson, D. K. *Data Reduction and Error Analysis for the Physical Sciences*, 3rd ed.; McGraw-Hill: New York, 2003.
(39) Flyvbjerg, H.; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461−466.
(40) Zwanzig, R.; Ailawadi, N. K. *Phys. Rev.* **1969**, *182*, 280−283.
(41) Wlodawer, A.; Walter, J.; Huber, R.; Sjölin, L. *J. Mol. Biol.* **1984**, *180*, 301−329.
(42) Xue, Y.; Ward, J. M.; Yuwen, T.; Podkorytov, I. S.; Skrynnikov, N. R. *J. Am. Chem. Soc.* **2012**, *134*, 2555−2562.
(43) Otting, G.; Liepinsh, E.; Wüthrich, K. *Biochemistry* **1993**, *32*, 3571−3582.
(44) Grey, M. J.; Wang, C.; Palmer, A. G. *J. Am. Chem. Soc.* **2003**, *125*, 14324−14335.
(45) Kaplan, E. L.; Meier, P. *J. Am. Stat. Assoc.* **1958**, *53*, 457−481.
(46) Lawson, C. L.; Hanson, R. J. *Solving Least Squares Problems*; Prentice-Hall: Englewood Cliffs, NJ, 1974.
(47) Whittall, K. P.; MacKay, A. L. *J. Magn. Reson.* **1989**, *84*, 134−152.