JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Efficient Explicit-Solvent Molecular Dynamics Simulations of Molecular Association Kinetics: Methane/Methane, Na^+/Cl^-, Methane/Benzene, and K^+/18-Crown-6 Ether

Matthew C. Zwier, Joseph W. Kaus, and Lillian T. Chong*

Department of Chemistry, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States

Supporting Information

ABSTRACT: Atomically detailed views of molecular recognition events are of great interest to a variety of research areas in biology and chemistry. Here, we apply the weighted ensemble path sampling approach to improve the efficiency of explicit solvent molecular dynamics (MD) simulations in sampling molecular association events between two methane molecules, Na^+ and Cl^- ions, methane and benzene, and the K^+ ion and 18-crown-6 ether. Relative to brute force simulation, we obtain efficiency gains of at least 300 and 1100-fold for the most challenging system, K^+/18-crown-6 ether, in terms of sampling the association rate constant k and distribution of times required to traverse transition paths, respectively. Our results indicate that weighted ensemble sampling is likely to allow for even greater efficiencies for more complex systems with higher barriers to molecular association.

1. INTRODUCTION

Proteins bind their partners in a highly specific manner. Understanding the mechanisms of these binding events is not only fundamentally interesting but could also impact fields such as protein engineering, host−guest chemistry, and drug discovery. Atomistic molecular dynamics (MD) simulations can potentially offer the most detailed views of molecular recognition events, especially when performed with explicit solvent. However, only up to a microsecond of simulation is practical on typical computing resources, while protein binding events require microseconds and beyond.[1] It is therefore computationally prohibitive to capture these events by sufficiently long "brute force" simulations. Fortunately, the long time scales required for protein binding events are not necessarily a result of the actual events taking a long time; instead, the events may be fast but infrequent, separated by long waiting times.

Path sampling approaches[2−10] aim to capture rare events by minimizing the simulation of long waiting times between events.[11] Weighted ensemble sampling[2] is one such approach which is rigorously correct for any type of stochastic simulation,[12] easily parallelized, and simultaneously provides both transition paths and their associated kinetics.[2] Weighted ensemble sampling has been applied to Brownian dynamics simulations of protein−protein binding,[2] protein−substrate binding,[13] protein folding,[14] Monte Carlo simulations of large-scale conformational transitions in the molecular switches calmodulin[15] and adenylate kinase,[16] and molecular dynamics simulations of alanine dipeptide in implicit solvent.[17]

We apply the weighted ensemble path sampling approach with explicit-solvent MD simulations. Our goal is to determine the efficiency of the weighted ensemble approach relative to brute force simulation in sampling molecular associations for a range of well-studied systems: methane/methane,[18−23] Na^+/Cl^-,[24−33] methane/benzene,[34,35] and K^+/18-crown-6 ether[36,37] (Figure 1). These systems were chosen because of their small size and relatively low barriers to association ($\sim 2k_BT$); combined, these features make feasible the simulation of association events by brute force, providing us with opportunities to evaluate not only the efficiency of the weighted ensemble approach but its validity as well.

2. THEORY

2.1. Overview of Weighted Ensemble Sampling. Weighted ensemble sampling uses "statistical ratcheting" to efficiently sample rare events using stochastic simulations.[2,11,15,17] To monitor the progress of these simulations toward the rare event of interest (here molecular association), a progress coordinate between the source (A, unbound) and destination (B, bound) states is defined by one or more order parameters; this progress coordinate is then divided into bins. A number of simulations are started in the unbound state A, which are then propagated for a fixed time τ. After this propagation time, if a simulation has progressed into a bin closer to the destination state B, its current state is used to start replicas of that simulation; these replicas diverge due to the stochastic nature of the underlying dynamics. Alternatively, if the simulation has regressed toward the source state A, it is effectively terminated. This resampling procedure[12] involving the replication of productive simulations and termination of unproductive simulations is repeated at fixed intervals (τ, 2τ, 3τ, and so on) until the desired number of rare events (crossings into state B) is sampled. Once a simulation reaches the destination state B, it is removed from the destination state B and "recycled" as a new simulation starting from the source state A. As this propagation and resampling procedure is repeated, the transition path ensemble—an ensemble of continuous trajectories between the source and destination states—is generated. As shown in Figure 2, some common history is shared among this

Received: November 2, 2010
Published: February 25, 2011

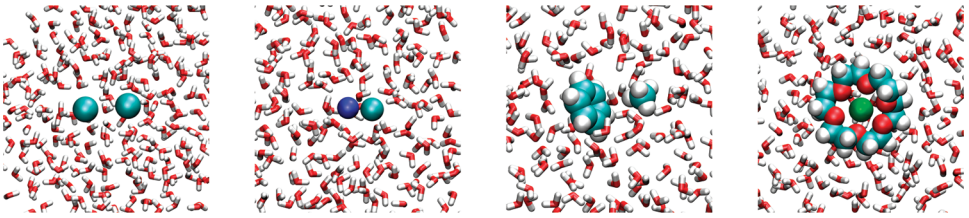

Figure 1. Molecular recognition systems of this study. From left to right, two methane molecules, Na^+/Cl^-, benzene and methane, and a K^+ ion with 18-crown-6 ether. All systems were immersed in explicit water molecules. (Prepared with VMD.[38])

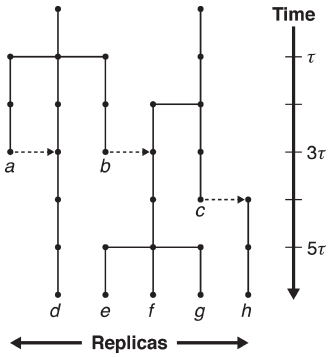

Figure 2. Schematic diagram of weighted ensemble molecular dynamics trajectories. Replication and termination of simulations occurs at intervals of τ, the propagation/resampling time. Trajectories *a* and *b* are terminated at $t = 3\tau$, and trajectory *c* reaches the destination state at $t = 4\tau$, at which time its statistical weight is assigned to a newly created replica which traces out trajectory *h*; the dotted arrows indicate a transfer of statistical weight but not history. Trajectories are replicated at $t = \tau$, $t = 2\tau$, and $t = 5\tau$. Note that trajectories *e*, *f*, and *g* share common history but are independent from trajectories *d* and *h*, which themselves are mutually independent.

ensemble of trajectories, and each trajectory has a maximum length τN_τ after N_τ iterations of propagation and resampling. When the trajectories are generated using molecular dynamics simulations, a stochastic thermostat is required to allow for divergence of trajectories after resampling.

To maintain correct statistics and kinetics of the transition paths, each simulation is assigned an appropriate statistical weight. When simulations are replicated, their statistical weights are split. When simulations are terminated for regressing toward the source state *A*, their statistical weights are merged into existing replicas, and when simulations are terminated for reaching the destination state *B*, their statistical weights are removed from the destination state *B* and assigned to newly created replicas in the initial state *A*.

2.2. Rate Constants. Weighted ensemble sampling yields kinetic information as a simulation progresses. After steady-state probability recycling is attained, the rate constant k is given by the average probability current I_B into the destination state *B*:[2,15,39]

$$k = \langle I_B \rangle \tag{1}$$

where the angle brackets indicate a time average. Because the recycling procedure described above eliminates all probability from the destination state *B* at each resampling, the probability current I_B may be approximated as

$$I_B \approx \frac{P_B(\tau N_\tau)}{\tau} \tag{2}$$

where τ is the weighted ensemble propagation/resampling time step and $P_B(\tau N_\tau)$ is probability contained in the destination state at time τN_τ (weighted ensemble iteration N_τ) immediately prior to recycling. Since $P_B(\tau N_\tau)$ must be monitored in order to ensure probability conservation during a weighted ensemble simulation, the rate constant k is obtained "for free."

The partially shared history of weighted ensemble trajectories results in highly correlated probability current measurements; that is, $I_B(\tau N_\tau)$ and $I_B(\tau[N_\tau - 1])$ are not statistically independent. The time average $\langle I_B \rangle$ may be computed in the usual way, but the associated confidence interval (encompassing the statistical error in the rate constant) must be computed with a method that accounts for the time correlation within I_B, such as Monte Carlo bootstrapping.[2,40,41]

On the other hand, the quantity accessible from brute force dynamics is not the probability current into the destination state but is rather a set of elapsed times between completed transition events. That is, brute force simulation does not yield the rate constant directly, but rather the first passage time distribution. For transitions dominated by a single time scale, this distribution is exponential, and the rate constant is simply the inverse of the mean first passage time $\langle t_{fp} \rangle$:[39]

$$k = \langle t_{fp} \rangle^{-1} \tag{3}$$

It should be noted that these two methods for determining the rate constant k are alternative mathematical descriptions of the same underlying physical principles (for an extensive discussion, see ref 39). Thus, rate constants obtained from brute force and weighted ensemble simulations may be directly compared, given that the same model was used for propagating dynamics in both cases and that the confidence interval for the rate constant is calculated correctly for the weighted ensemble simulation.

2.3. Transition Event Durations. If a system exhibits rare but fast events, then the transition event duration t_{ed}—the amount of time it takes a transition to complete once it starts—is much less than the mean first passage time $\langle t_{fp} \rangle$ (which includes the waiting time between rare events):

$$t_{ed} \ll \langle t_{fp} \rangle$$

The probability distribution of t_{ed}, $F(t_{ed})$, is at least approximately an indicator of the extent of sampling of the transition pathways. Distinct pathways will have associated characteristic transition durations,[42] and as transition pathways are sampled, $F(t_{ed})$ will become better resolved. Thus, the self-convergence of $F(t_{ed})$ is a strong indicator that the transition path ensemble has been adequately explored.

The transition event duration distribution $F(t_{ed})$ is built up directly from simulation trajectories simply by noting the time elapsed between exiting the source state *A* and entering the destination state *B*. In the brute force case, a set of event

durations is transformed directly into a cumulative distribution function in the usual manner (by counting the number of t_{ed} less than a specified value):

$$F(t_{ed}) = \frac{1}{N}\sum_i h(t_{ed(i)}) \quad (4)$$

where i indexes transitions, N is the number of transition events observed, $t_{ed(i)}$ is the duration of transition event i, and h is an indicator function satisfying

$$h(t_{ed(i)}) = \begin{cases} 1 & \text{if } t_{ed(i)} \leq t_{ed} \\ 0 & \text{otherwise} \end{cases}$$

Weighted ensemble simulations, on the other hand, yield not the set of event durations $\{t_{ed}\}$ but a set $\{(t_{ed},w)\}$ of transition event durations and corresponding terminal statistical weights. These terminal weights partially encode the probability of arriving at the final state, and so a weighted variation of eq 4 must be used:

$$F(t_{ed}) = \frac{\sum_i w_i h(t_{ed(i)})}{\sum_i w_i} \quad (5)$$

There are several advantages to describing the transition event duration distribution as an (empirical) cumulative distribution function. First, rigorous confidence bands may be assigned to empirical distribution functions,[43,44] allowing one to assign error bars to the entire t_{ed} distribution and facilitating the comparison of simulation results. Second, the number of points N_e in a realization of $F(t_{ed})$ is equal to the number of unique transition event durations sampled and, as such, can be considered a statistical sample size for the purposes of quantifying sampling, even in the weighted ensemble case. For this same reason, eq 5, despite being cast in a weighted form, describes a formal empirical distribution function and is therefore an unbiased estimator of the true cumulative distribution function.[44]

2.4. Relative Efficiency of Weighted Ensemble Simulations. Any meaningful metric for comparing the relative efficiencies of weighted ensemble and brute force simulations must account for not only the computational expense of obtaining an estimate on a quantity such as the rate constant, but also the uncertainty of that estimate. In other words, an efficiency metric must take error bars into account. For a given quantity like the reaction rate k, we define the efficiency of weighted ensemble sampling relative to brute force as

$$S = \frac{t_{(WE)}}{t_{eff}} \quad (6)$$

where $t_{(WE)}$ is the aggregate weighted ensemble simulation time (not overcounting shared history) and t_{eff} is the effective amount of brute force simulation time that would be required to obtain an estimate with the same size error bar as that obtained from a weighted ensemble simulation. Consideration of the error structure of brute force simulations and application of eq 6 gives the following efficiency metrics S_k and S_{ed} for sampling of the association rate constant k and t_{ed} distribution, respectively:

$$S_k = \frac{t_{(BF)}}{t_{(WE)}}\left(\frac{\Delta k^*_{(BF)}}{\Delta k^*_{(WE)}}\right)^2 \quad (7)$$

$$S_{ed} = \frac{t_{(BF)}}{t_{(WE)}}\left(\frac{N_{e(WE)}}{N_{e(BF)}}\right) \quad (8)$$

where t represents total simulation time, Δk^* is the width of the 95% confidence interval on the rate constant k relative to the time average $\langle k \rangle$, and N_e is the number of unique time values in the empirical distribution function $F(t_{ed})$; the subscripts (BF) and (WE) represent values from brute force and weighted ensemble simulations, respectively. Detailed derivations of eqs 7 and 8 are provided in the Supporting Information.

3. METHODS

3.1. Model Systems. Four systems were used to test the feasibility of using weighted ensemble sampling with explicit-solvent MD simulations to study molecular association events. These systems all possess simple, one-dimensional progress coordinates by which it is possible to unambiguously define "how close to binding" a simulation is at any point in time. All systems were immersed in boxes of explicit water molecules. The model systems in order of progressively more challenging features are described below.

Methane/Methane. This system is a simple example of a hydrophobic interaction. The natural progress coordinate of this system is simply the center-to-center distance between the two methane molecules.

Na^+/Cl^-. This system is a simple example of an electrostatic interaction. The natural progress coordinate of this system is the center-to-center distance between the two ions.

Methane/Benzene. Like the methane/methane system, methane/benzene is a model of hydrophobic interactions, but unlike the previous two systems, it does not exhibit an effective spherical symmetry. However, our brute force simulations of this system revealed that the condensed-phase bound state involves precession of the methane molecule about the surface of the benzene ring. Therefore, despite the broken spherical symmetry, the natural progress coordinate for this system is effectively one-dimensional and was taken to be the distance between the methane carbon and the center of mass of the benzene carbon atoms.

K^+/18-crown-6 ether. This system is a simple example of the binding of a (trivially) rigid substrate (K^+) by a flexible partner (18-crown-6 ether). Like methane/benzene, this system does not exhibit effective spherical symmetry. However, both simulation[36,37] and X-ray crystallography[45] have indicated that the bound structure consists of the K^+ ion coplanar with the crown ether oxygen atoms. The natural progress coordinate for this system is therefore the distance between the K^+ ion and the center of mass of the ether oxygen atoms.

3.2. Simulation Details. Both brute force and weighted ensemble simulations were performed using the GROMACS 4.0.5 software package.[46] Production dynamics (both brute force and weighted ensemble) were propagated in the canonical (NVT) ensemble at 300 K using a Langevin thermostat[47] (coupling time 1 ps). Van der Waals interactions were switched off smoothly between 8 and 9 Å; to account for the truncation of the van der Waals interactions, a long-range analytical dispersion correction[48] was applied to energy and pressure. Real-space electrostatic interactions were truncated at 10 Å. Long range electrostatic interactions were calculated using particle mesh

Ewald[49] (PME) summation. Bonds to hydrogen atoms were constrained to their equilibrium lengths using LINCS,[50] permitting a 2 fs integration time step.

Each model system was constructed in its unbound state and solvated in a dodecahedral periodic box with a minimum 12 Å clearance between the solutes and the box walls. Following a 1000-step steepest-descent energy minimization, each system was subjected to 20 ps of NVT thermal equilibration followed by 1 ns of constant-pressure (NPT) density equilibration using a weak isotropic Berendsen barostat[51] (reference pressure 1 bar, coupling time 5 ps, and compressibility 4.5×10^{-5} bar^{-1}). In both equilibration stages, all heavy atoms were restrained using a harmonic potential. The resulting equilibrated systems were used as starting points for both brute force and weighted ensemble MD simulations. The initial pair separations were 10, 10, 17, and 15 Å for methane/methane, Na^+/Cl^-, methane/benzene, and K^+/18-crown-6 ether, respectively. The GROMOS 45A3 united-atom force field[52] and SPC/E[53] water model were used for methane/methane and Na^+/Cl^-, while the OPLS-AA/L force field[54] and the TIP3P[55] water model were used for methane/benzene and K^+/18-crown-6 ether. Atom type assignments for K^+/18-crown-6 ether ether are provided in the Supporting Information (Figure S1).

3.3. Brute Force Dynamics Propagation. Brute force simulations for all model systems were started from the end points of their respective second-stage (density) equilibration runs. Each simulation was continued until a sufficient number of transition events was observed, with solute positions recorded every 10 fs. The methane/methane and methane/benzene systems were both run as single 1 μs trajectories. Na^+/Cl^- and K^+/18-crown-6 ether required multiple independent trajectories to observe a sufficient number of transition events. A total of 10 independent 1 μs trajectories were run for Na^+/Cl^-, and 100 independent 100 ns trajectories were run for K^+/18-crown-6 ether.

3.4. Determination of Bound and Unbound States. The analysis of brute force trajectories and the construction of weighted ensemble simulations require unambiguous definitions of bound and unbound states for each system. Because all four model systems possess one-dimensional progress coordinates, the same protocol for determining these states was applied to all four model systems. Pairwise condensed-phase interactions can be described by the potential of mean force (PMF) $u(r)$, the free energy of the system as a function of pair separation r.[56] Taking the zero of energy to be the noninteracting limit, for constant-volume systems the $u(r)$ is given by the following:[23]

$$u(r)/k_BT = -\left(\ln\frac{P(r)}{r^2} - \ln\frac{P(r_0)}{r_0^2}\right) \qquad (9)$$

where $P(r)$ is the probability of observing the system at a pair separation r, r_0 is the shortest distance at which the pair is effectively noninteracting ($du/dr \approx 0$ for all $r > r_0$), and the factors of r^2 arise from the transformation between the Cartesian coordinates of the MD simulation and the spherical polar coordinates in which $u(r)$ is expressed. For each model system, the PMF $u(r)$ was determined using eq 9 with pairwise distance probabilities $P(r)$ taken from the brute force trajectories. The unbound state A was defined as $A = \{r: r \geq r_0\}$, where (as above) r_0 is the shortest distance at which the pair is effectively noninteracting. This definition ensures that binding events observed in brute force simulations are very nearly statistically independent. The bound state B was readily identified as being near the global minimum of $u(r)$ and defined as $B = \{r: r < r_B\}$, where r_B is the separation at which the global minimum well of $u(r)$ becomes concave up; that is, B is the basin of attraction of the global minimum of $u(r)$. The remainder of progress coordinate space defines a transition region $T = \{r: r_B \leq r < r_0\}$ wherein the partners are interacting but not definitively bound. PMF curves for each system are provided in Figures S3−S6 of the Supporting Information.

3.5. Determination of Weighted Ensemble Simulation Parameters. In addition to definitions of bound and unbound states, a weighted ensemble simulation requires selection of optimal bin sizes, numbers of replicas per bin, and propagation/resampling interval τ. In making these selections, the extent of sampling should be maximized (generally meaning more bins and more replicas per bin) while minimizing the overall computational cost (generally meaning fewer bins and fewer replicas per bin).

For all four model systems, the potential of mean force was used to determine a bin spacing aimed at maximizing the "ratcheting" effect of the weighted ensemble approach. Where the PMF was changing rapidly with respect to pair separation, bin boundaries were chosen such that the crossing of a bin does not require climbing more than $\sim k_BT$ in energy as indicated by the appropriate PMF. This ensures that the system can move about the progress coordinate with relative ease. Conversely, in the region where the PMF is slowly varying, a constant spacing of bins was adopted. The propagation period τ was then chosen so that the RMS change in pair separation over a time τ was approximately equal to the width of the bins in the slowly varying region of the PMF. This resulted in bins of width ∼0.1−1.0 Å. Initial tests indicated that 50 replicas per bin yielded sufficiently precise values for the rate constant k at a reasonable computational cost, so this value was used for all four model systems. Detailed listings of the resulting bin boundaries are provided in Figures S3−S6 (Supporting Information), and the remaining weighted ensemble sampling parameters are summarized in Table S1 (Supporting Information).

3.6. Weighted Ensemble Dynamics Propagation. Weighted ensemble dynamics runs used exactly the same simulation parameters (force field, thermostat parameters, box volume, etc.) as those of the corresponding brute force simulations. As with the brute force simulations, the initial atomic coordinates and velocities were taken from the end of the equilibration phase for each model system. The weighted ensemble sampling algorithm was implemented in an in-house computer code as described above. Replicas were propagated in parallel on 32−96 CPU cores, requiring a few days to simulate each model system. Both the rate constant k and the transition event duration distribution $F(t_{ed})$ were monitored every 50 or 100 τ, and the weighted ensemble simulation was terminated when k was constant within uncertainty and $F(t_{ed})$ had converged to within 95% confidence and remained at that level, as determined by a two-sided Kolmogorov−Smirnov test[44] (a standard test of the statistical equivalence of two empirical distribution functions). Though resampling was performed with a period of τ, all analysis of the simulations was conducted at a time resolution of 10 fs (the period with which solute positions were recorded during the underlying dynamics simulations). The resulting aggregate simulation times for each system are presented in Table S2 (Supporting Information).

Table 1. Brute Force (BF) and Weighted Ensemble (WE) Aggregate Simulation Times t, Rate Constants (k), and Relative Sampling Efficiencies (S_k) for the Four Model Systems[a]

system	t_{BF}	t_{WE}	k_{BF} (ps^{-1})	k_{WE} (ps^{-1})	S_k
methane/methane	1 μs	299 ns	$1.91 \pm 0.10 \times 10^{-3}$	$1.61 \pm 0.06 \times 10^{-3}$	7.0
Na^+/Cl^-	10 μs	3.86 μs	$1.86 \pm 0.09 \times 10^{-4}$	$1.82 \pm 0.11 \times 10^{-4}$	1.4
methane/benzene	1 μs	369 ns	$8.6 \pm 0.7 \times 10^{-4}$	$7.7 \pm 0.3 \times 10^{-4}$	8.7
K^+/18-crown-6	10 μs	322 ns	$2.1 \pm 0.3 \times 10^{-5}$	$4.8 \pm 0.2 \times 10^{-5}$	300

[a] Aggregate simulation times correspond to the combined length of all trajectories (either brute force or weighted ensemble) for each system, without overcounting common history in the case of weighted ensemble simulations. Uncertainties on the rate constants represent 95% confidence intervals. Relative efficiencies were calculated using eq 7.

Table 2. Ratios of Rate Constants k and Average Waiting Times $\langle t_w \rangle$ for Brute Force (BF) and Weighted Ensemble (WE) Simulations

system	$k_{(WE)}/k_{(BF)}$	$\langle t_w \rangle_{(BF)}/\langle t_w \rangle_{(WE)}$
methane/methane	0.842	0.841
Na^+/Cl^-	0.977	0.977
methane/benzene	0.827	0.822
K^+/18-crown-6	1.93	1.94

4. RESULTS AND DISCUSSION

The purpose of this study was to determine the efficiency of weighted ensemble sampling relative to brute force sampling for association events in four molecular recognition systems. As described above, both the association rate constant k and the transition event duration distribution $P(t_{ed})$ can be used to quantify sampling of the transition path ensemble. We compare the efficiency and accuracy of weighted ensemble simulations relative to brute force simulations in terms of both rate constants and transition event distributions.

4.1. Rate Constants. The rate constant (k) values for brute force and weighted ensemble simulations were separately converged to within statistical uncertainty. As shown in Table 1, the weighted ensemble simulations are in qualitative agreement with brute force simulations for all systems; quantitative agreement was achieved for Na^+/Cl^- and methane/benzene. The relative efficiency S_k of weighted ensemble sampling of the rate constant was modest (1.4-fold) for Na^+/Cl^-, greater than 5-fold for the diffusive systems (methane/methane and methane/benzene), and 300-fold for the most complex system, K^+/18-crown-6 ether.

It is not surprising that the rate constant obtained by weighted ensemble sampling for K^+/18-crown-6 ether does not agree with the brute force simulation, as the brute force $P(t_{ed})$ did not converge; it is less clear why the rate constants for methane/methane are not in agreement. One possibility is that either the brute force or the weighted ensemble simulation did not sample the full set of waiting times between rare events. The waiting time t_w between subsequent $A \rightarrow B$ transition events relates the first passage time t_{fp} and the transition event duration t_{ed} according to

$$t_{fp} = t_{ed} + t_w$$

In all cases (including that in which t_{ed} and t_w are not statistically independent):

$$\langle t_{fp} \rangle = \langle t_{ed} \rangle + \langle t_w \rangle$$

where the angle brackets denote the expectation (mean) value. Since $\langle t_{ed} \rangle \ll \langle t_{fp} \rangle$ for all four systems considered here, the discrepancy between brute force and weighted ensemble simulations in mean waiting time $\langle t_w \rangle$ accounts almost completely for the discrepancy in rate constants between simulation techniques (see Table 2). It is likely that the overestimated brute force waiting time for K^+/18-crown-6 ether is due to poor convergence of the brute force simulation. Similarly, it seems likely that the methane/methane brute force simulation underestimated t_w for that system. In both of these cases, the efficiencies presented in Table 1 represent lower bounds, as they assume complete convergence of the brute force simulations.

Implicit in the foregoing analysis is the assumption that the first passage time distribution $P(t_{fp})$ obtained from the brute force simulation is exponential, as would be the case in a system possessing (effectively) a single barrier of constant height; that is,

$$P(t_{fp}) = 1 - \exp(-kt_{fp}) \quad (10)$$

where k is the rate constant. In this case, the rate constant k is equal to the inverse mean first passage time [cf. eq 3]. If the first passage time distribution $P(t_{fp})$ is not exponential, then the inverse mean first passage time is at best an approximation of the true rate constant; conversely, the weighted ensemble approach samples k directly, and so it can be expected to recover the correct rate constant (within the bounds of statistical uncertainty) regardless of whether the underlying physical mechanisms lead to an exponential first passage time distribution. For three of the four model systems (Na^+/Cl^-, methane/benzene, and K^+/18-crown-6), the first passage time distributions obtained from brute force simulations conform to eq 10 to within 95% confidence (see Figure S2, Supporting Information). For methane/methane, however, the first passage time distribution deviates from the expected exponential distribution for $t_{fp} \lesssim 300$ ps. This offers an alternative explanation for why the rate constant values obtained for methane/methane differ between brute force and weighted ensemble simulations: because the first passage time distribution $P(t_{fp})$ is not exponential, the rate constant k obtained from the brute force first passage time distribution as $\langle t_{fp} \rangle^{-1}$ may in fact be inaccurate.

4.2. Transition Event Duration Distributions. In general, the weighted ensemble simulations were as good or better than brute force simulations in generating well-resolved transition event duration distributions $P(t_{ed})$. As shown in Figure 3, $P(t_{ed})$ was well-resolved by both brute force and weighted ensemble simulations for all systems except K^+/18-crown-6 ether, for which brute force sampling was not capable of providing a converged $P(t_{ed})$ distribution. The resolution of distributions from weighted ensemble simulations far exceeds that of distributions obtained from brute force simulations, as demonstrated in the increased number N_e of unique transition durations sampled (see Table 3). Further, pathways generated by weighted ensemble sampling and having different transition event durations were

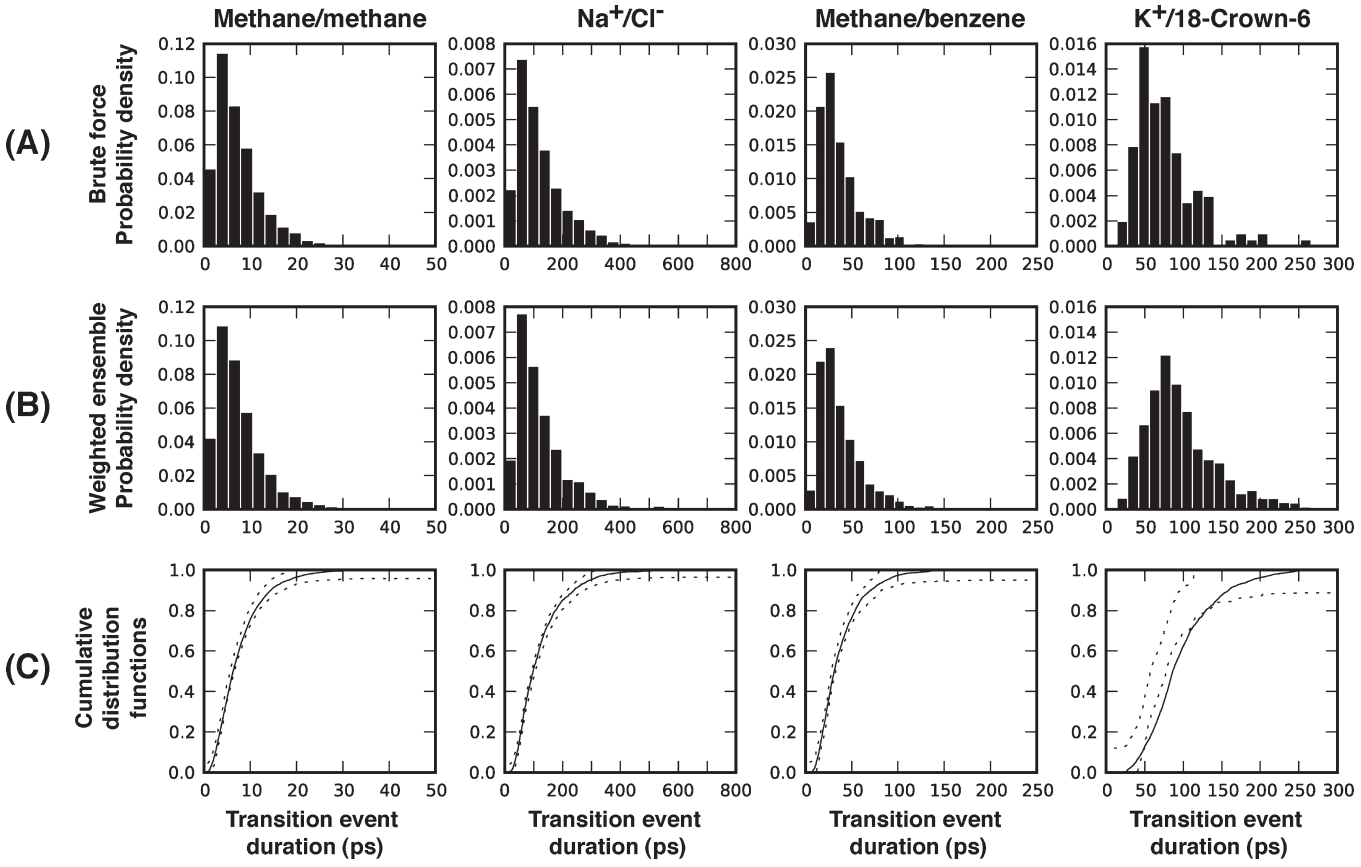

Figure 3. Transition event duration distributions obtained from (A) brute force and (B) weighted ensemble simulations. The cumulative distribution function (CDF) of the transition event duration probability for each model system is shown in (C). The brute-force CDF is plotted as a 95% confidence interval with dotted lines, and the solid line is the CDF obtained from the weighted ensemble simulation.

Table 3. Number of Unique Transition Durations N_e and Relative Efficiency S_{ed} of Sampling of the Transition Event Duration Distribution for Brute Force (BF) and Weighted Ensemble (WE) Simulations[a]

system	$N_{e(BF)}$	$N_{e(WE)}$	S_{ed}
methane/methane	1021	2304	7.5
Na^+/Cl^-	1415	8780	16
methane/benzene	750	5485	20
K^+/18-crown-6	145	5007	1100

[a] Relative efficiency was calculated using eq 8.

indeed noticeably different from each other (see Supporting Information, section S.1, Figures S1 and S2, and Movies S1 and S2). These are strong indications that the weighted ensemble algorithm effectively enhances sampling of the transition path ensemble. The relative efficiency S_{ed} of sampling $F(t_{ed})$ increased with the complexity of the molecular recognition system, ranging from 1 to 3 orders of magnitude. The 1100-fold relative efficiency of weighted ensemble sampling for K^+/18-crown-6 ether is a conservative estimate, as the referenced brute force simulation had not even reached convergence with respect to $F(t_{ed})$.

As shown in Tables 1 and 3, $S_k < S_{ed}$ in all four cases. This is partly a consequence of our definitions of the efficiency metrics S_k and S_{ed} (see above and Supporting Information), but it also reflects that the rate constant k is generally more difficult to sample than the set of transition event durations $\{t_{ed}\}$. In particular, convergence of the rate constant k requires sampling of *all* important pathways as well as a steady state flow of probability through them.

4.3. How Much Sampling Is Required? As evident for K^+/18-crown-6 ether, the most complex system of this study, it is not always possible to obtain converged brute force simulations of molecular association events. In such cases, how does one know if the weighted ensemble approach has achieved sufficient sampling? One can, at least, gauge the self-convergence of the association rate constants k and the transition event duration distributions $F(t_{ed})$ obtained from the weighted ensemble simulations. However, self-convergence of these metrics does not guarantee that the simulation has converged to the true value of k or $F(t_{ed})$.

As an illustration, consider the convergence of $F(t_{ed})$, the probability distribution of the event duration times Ft_{ed}. Even if two transition event distributions $F_{\tau(1)}(t_{ed})$ and $F_{\tau(2)}(t_{ed})$ obtained by time points $N_{\tau(1)}$ and $N_{\tau(2)} > N_{\tau(1)}$ in a weighted ensemble simulation are statistically equivalent, this does not necessarily indicate asymptotic convergence on the true transition event duration distribution. Because a weighted ensemble simulation of length iterations contains only trajectories of maximum length τN_τ, then the statistical equivalence of $F_{\tau(1)}(t_{ed})$ and $F_{\tau(2)}(t_{ed})$ does not indicate that the entire event duration distribution has been adequately sampled, merely that all pathways taking time $t \leq \tau N_{\tau(1)}$ to traverse have been adequately sampled. Thus, for a weighted ensemble simulation of length

τN_τ, one must ultimately decide whether data obtained for time scales less than τN_τ are sufficient to provide insights into the systems under study.

4.4. How Does One Choose Optimal Weighted Ensemble Parameters? Efficient use of weighted ensemble sampling involves finding the optimal balance between computational expense and level of sampling. A poor choice of progress coordinate bins can easily lead to oversampling relatively unimportant regions of phase space. A large number of replicas not only aids rapid exploration of phase space but also determines the precision of probability current value and thus kinetic information; however, the total computational cost of weighted ensemble scales approximately linearly with the maximum number of system replicas. A short propagation/resampling period τ allows many opportunities for replicas to split and explore newly visited regions of phase space and for replicas to merge to avoid oversampling regions of phase space but ultimately may not allow sufficient divergence of trajectories to allow for efficient exploration of phase space.

Integral to the construction of a weighted ensemble simulation is the choice of a progress coordinate that is sufficiently sensitive to quantify "how far along" the reaction is. Any number of relatively low-cost enhanced-sampling or energy landscape smoothing techniques[57−59] might be employed to guide the choice of a progress coordinate, including metadynamics;[60,61] targeted,[62] steered,[63] or accelerated[64] molecular dynamics; or the recently developed orthogonal space random walk method.[65] A number of short brute force simulations may be required to determine the average time evolution of the progress coordinate, which in turn determines the most efficient choices of bin spacing and the propagation/reweighting period τ. Finally, it may be necessary to adjust these parameters "on the fly" during a simulation, especially for large systems with complex, rough energy landscapes (i.e., proteins) where long-lived intermediate states may be encountered in the course of a simulation.

The complexities and advantages of actively adjusting the numbers of bins, their boundaries, and the number of replicas in each bin have been discussed in detail;[2] such schemes could be used to detect replicas that "stall" in certain progress coordinate bins and adjust the weighted ensemble simulation to compensate. These schemes would not be able to cope effectively with systems possessing intermediate states with lifetimes comparable to the mean first passage time; such systems do not exhibit the separation of time scales which weighted ensemble sampling is designed to exploit. However, using ideas developed from nonequilibrium umbrella sampling, it is possible to reweight phase space density analytically in order to accelerate the attainment of steady-state probability recycling;[17] this would in turn accelerate the determination of the rate constant in systems with $t_{ed} \approx t_{fp}$ at the possible expense of efficient sampling of the transition path ensemble.

Finally, it should be noted that the weighted ensemble approach is but one instance of a class of "interface-based" enhanced sampling techniques which share a number of strengths and potential weaknesses;[11,66,67] other such techniques include transition interface sampling (TIS) and variants,[5,6,10] forward flux sampling (FFS),[8,9] and milestoning.[7] All of the methods in this class are rare event sampling methods that divide phase space along distinct interfaces, and each method is capable of providing realistic kinetic rates. Provided a well-chosen progress coordinate, these methods are equivalent in principle with respect to the information which can be obtained from them and the efficiency with which that information is obtained, at least for equilibrium systems. Among these methods, however, the weighted ensemble approach is uniquely flexible; in particular, sampling can be maximized while minimizing computational cost both by dividing phase space according to arbitrary boundaries in any number of dimensions and by adjusting the level of sampling within each region (by adjusting the number of simulation replicas within a bin). The cost of this flexibility, however, is the complexity of determining efficient choices for parameters such as the progress coordinate, bin boundaries, and the number of replicas per bin. In situations where a reasonable progress coordinate cannot be determined, a method not dependent on a progress coordinate (such as transition path sampling[3,68,69] or a recently developed variation of milestoning[70]) may be necessary. Similarly, if efficient choices for simulation parameters (such as bin boundaries and the number of replicas per bin) cannot be made in advance and adjustment of these parameters during a simulation is impractical, then a method like FFS (for which analytical expressions for efficiency as a function of simulation parameters exist[71,72]) may be a better choice.

4.5. Why Are Efficiencies What They Are? The efficiency of a weighted ensemble simulation is largely determined by weighted ensemble simulation parameters, particularly the propagation/resampling period τ, the choice of progress coordinate(s), and the locations of bin boundaries.[42] For some systems, brute force simulation is already highly efficient at sampling the molecular association events; this is confirmed by the modestly increased weighted ensemble sampling efficiencies (S_k and S_{ed}) for methane/methane, Na^+/Cl^-, and methane/benzene. However, the fact that the weighted ensemble approach increases rather than decreases efficiency indicates that, even in such cases, the weighted ensemble technique is capable of accelerating sampling of both k and $F(t_{ed})$. On the other hand, the very high relative efficiency of sampling in K^+/18-crown-6 ether is particularly encouraging. Despite the small size of the system, brute force MD was incapable of effective sampling of rate constants and transition event duration distributions for K^+/18-crown-6 ether, almost certainly due to the high (approximately 14 k_BT, 8.3 kcal/mol) barrier to dissociation. Weighted ensemble sampling was able to obtain self-converged values of both the rate constant k and the transition event duration distribution $F(t_{ed})$. This is primarily because probability recycling completely circumvents the necessity to climb the 14 k_BT dissociation barrier in order to observe another binding event.

These results point encouragingly to the ability to simulate protein−protein binding events with weighted ensemble molecular dynamics. With well-chosen bin boundaries, the weighted ensemble technique should increase sampling efficiency exponentially with increasing barrier heights. This is because placing bin boundaries sufficiently close to each other effectively linearizes the probability of crossing a number of bins in succession, rather than surmounting a barrier in one step with a probability which decreases exponentially with barrier height.[73] As a concrete example, the barrier to association in a diffusion-limited protein−protein system is approximately 10 k_BT (roughly five times that of the model systems). If this exponential efficiency scaling holds, then one can expect about 20 000-fold improvement in sampling for such a system. In other words, if a given computational resource is otherwise capable of generating 500 ps per calendar day (a substantial but accessible level of computational power), this efficiency gain corresponds to reaching a time scale of about 1 ms in 100 days, compared to the 50 ns that would

otherwise be possible in the same amount of time. However, since protein−protein binding pathways involve significant metastable intermediate states (e.g., encounter complexes[74]), it is possible for a simulation to "stall" in such a state. As discussed above, several techniques exist which may partially ameliorate this difficulty, but in the end, a number of simulations connecting the intermediate states may be necessary to fully explore binding events in such systems.

5. CONCLUSIONS

We have applied the weighted ensemble path sampling approach to molecular dynamics simulations in explicit solvent, enabling the detailed sampling of rare molecular association events. We have compared the efficiency of weighted ensemble sampling relative to brute force sampling in simulating association events of methane/methane, Na^+/Cl^-, methane/benzene, and K^+/18-crown-6 ether. Relative to brute force simulation, weighted ensemble sampling of these four systems confirms that the weighted ensemble approach reproduces or even improves sampling of both the rate constant k and the distribution of transition event durations. This improvement is on the order of 300- and 1100-fold, respectively, for a system exhibiting significant conformational flexibility (K^+ binding with 18-crown-6 ether). We expect efficiency gains to grow with increasing barriers to association. However, the existence of significant metastable intermediate states may hinder sampling in such systems, requiring the use of various enhancements to the weighted ensemble method in order to explore binding events in such systems. Nonetheless, these results indicate that weighted ensemble sampling in conjunction with MD simulations is likely to allow for the effective determination of transition paths and rate constants for protein binding events.

■ ASSOCIATED CONTENT

Supporting Information. OPLS/AA atom type assignments for methane/benzene and K^+/18-crown-6 ether systems, brute force first passage time distributions and potential of mean force curves for model systems, discussion and movies of binding pathways for the K^+/18-crown-6 ether system, bin boundaries for weighted ensemble simulations, and derivations of the efficiency metrics S_k and S_{ed}. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: ltchong@pitt.edu.

■ ACKNOWLEDGMENT

We thank Dan Zuckerman and Divesh Bhatt (U. Pitt. Dept. of Computational Biology), Bin Zhang (U. Michigan Dept. of Chemistry), Michael Grabe and Josh Adelman (U. Pitt. Dept. of Biological Sciences), Gary Huber (UCSD Dept. of Bioengineering), and Karen Zwier and Jonathan Livengood (U. Pitt. Dept. of History and Philosophy of Science) for helpful discussion; we also thank Xianghong Qi for initial efforts. This work was supported by NSF CAREER award MCB-0845216 to L.T.C., a University of Pittsburgh Arts & Sciences Fellowship to M.C.Z., and a University of Pittsburgh Brackenridge Fellowship (underwritten by the United States Steel Foundation) to J.W.K.

■ REFERENCES

(1) Henzler-Wildman, K. A.; Kern, D. *Nature* **2007**, *450*, 964.
(2) Huber, G. A.; Kim, S. *Biophys. J.* **1996**, *70*, 97.
(3) Dellago, C.; Bolhuis, P. G.; Csajka, F.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964.
(4) Zuckerman, D. M.; Woolf, T. B. *Phys. Rev. E* **2000**, *63*, 1.
(5) van Erp, T. S.; Moroni, D.; Bolhuis, P. G. *J. Chem. Phys.* **2003**, *118*, 7762.
(6) Moroni, D.; Bolhuis, P. G.; van Erp, T. S. *J. Chem. Phys.* **2004**, *120*, 4055.
(7) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.
(8) Allen, R. J.; Warren, P.; ten Wolde, P. R. *Phys. Rev. Lett.* **2005**, *94*, 018104.
(9) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 024102.
(10) van Erp, T. S. *Phys. Rev. Lett.* **2007**, *98*, 1.
(11) Zwier, M. C.; Chong, L. T. *Curr. Opin. Pharm.* **2010**, *10*, 745.
(12) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *132*, 054107.
(13) Rojnuckarin, A.; Livesay, D. R.; Subramaniam, S. *Biophys. J.* **2000**, *79*, 686.
(14) Rojnuckarin, A.; Kim, S.; Subramaniam, S. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 4288.
(15) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 18043.
(16) Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 3527.
(17) Bhatt, D.; Zhang, B. W.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *133*, 014110.
(18) Dang, L. X. *J. Chem. Phys.* **1994**, *100*, 9032.
(19) Meng, E. C.; Kollman, P. A. *J. Phys. Chem.* **1996**, *100*, 11460.
(20) Oostenbrink, C.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2005**, *7*, 53.
(21) Trzesniak, D.; van Gunsteren, W. F. *Chem. Phys.* **2006**, *330*, 410.
(22) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2007**, *129*, 14887.
(23) Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, W. F. *Chemphyschem* **2007**, *8*, 162.
(24) Belch, A. C.; Berkowitz, M. L.; McCammon, J. A. *J. Am. Chem. Soc.* **1986**, *108*, 1755.
(25) Dang, L. X.; Rice, J. E.; Kollman, P. A. *J. Chem. Phys.* **1990**, *93*, 7528.
(26) Guàrdia, E.; Rey, R.; Padró, J. A. *Chem. Phys.* **1991**, *155*, 187.
(27) Hummer, G.; Soumpasis, D.; Neumann, M. *Mol. Phys.* **1992**, *77*, 769.
(28) Pratt, L. R.; Hummer, G.; Garcia, A. E. *Biophys. Chem.* **1994**, *51*, 147.
(29) Koneshan, S.; Rasaiah, J. C. *J. Chem. Phys.* **2000**, *113*, 8125.
(30) Patra, M.; Karttunen, M. *J. Comput. Chem.* **2004**, *25*, 678.
(31) Baumketner, A. *J. Chem. Phys.* **2009**, *130*, 104106.
(32) Fennell, C. J.; Bizjak, A.; Vlachy, V.; Dill, K. A. *J. Phys. Chem. B* **2009**, *113*, 6782.
(33) Timko, J.; Bucher, D.; Kuyucak, S. *J. Chem. Phys.* **2010**, *132*, 114510.
(34) Tsuzuki, S.; Honda, K.; Uchimaru, T.; Mikami, M.; Tanabe, K. *J. Am. Chem. Soc.* **2000**, *122*, 3746.
(35) Ringer, A. L.; Figgs, M. S.; Sinnokrot, M. O.; Sherrill, C. D. *J. Phys. Chem. A* **2006**, *110*, 10822.
(36) Dang, L. X.; Kollman, P. A. *J. Am. Chem. Soc.* **1990**, *112*, 5716.
(37) Troxler, L.; Wipff, G. *J. Am. Chem. Soc.* **1994**, *116*, 1468.
(38) Humphrey, W. *J. Mol. Graphics* **1996**, *14*, 33.
(39) Hänggi, P.; Talkner, P.; Borkovec, M. *Rev. Mod. Phys.* **1990**, *62*, 251.
(40) Efron, B. Y. B.; Tibshirani, R. *Stat. Sci.* **1986**, *1*, 54.
(41) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Cambridge University Press: Cambridge, England, 1992.

(42) Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2007**, *126*, 074504.
(43) Kolmogoroff, A. *Ann. Math. Stat.* **1941**, *12*, 461.
(44) Kvam, P. H.; Vidakovic, B. *Nonparametric Statistics with Applications to Science and Engineering*; John Wiley & Sons: Hoboken, NJ, 2007.
(45) Cambillau, C.; Bram, G.; Corset, J.; Riche, C.; Pascard-Billy, C. *Tetrahedron* **1978**, *34*, 2675.
(46) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
(47) Adelman, S.; Doll, J. *J. Chem. Phys.* **1976**, *64*, 2375.
(48) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.
(49) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.
(50) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.
(51) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.
(52) Schuler, L.; Daura, X.; van Gunsteren, W. F. *J. Comput. Chem.* **2001**, *22*, 1205.
(53) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
(54) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.
(55) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(56) Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press: New York, NY, 1987.
(57) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151.
(58) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589.
(59) Lei, H.; Duan, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 187.
(60) Huber, T.; Torda, a. E.; van Gunsteren, W. F. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695.
(61) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562.
(62) Schlitter, J.; Engels, M.; Krüger, P. *J. Mol. Graphics* **1994**, *12*, 84.
(63) Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K. *Biophys. J.* **1997**, *72*, 1568.
(64) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919.
(65) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227.
(66) Allen, R. J.; Valeriani, C.; Rein ten Wolde, P. *J. Phys.: Condens. Matter* **2009**, *21*, 463102.
(67) Escobedo, F. A.; Borrero, E. E.; Araque, J. C. *J. Phys.: Condens. Matter* **2009**, *21*, 333101.
(68) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291.
(69) Grünwald, M.; Dellago, C.; Geissler, P. L. *J. Chem. Phys.* **2008**, *129*, 194101.
(70) Májek, P.; Elber, R. *J. Chem. Theory Comput.* **2010**, *6*, 1805.
(71) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 194111.
(72) Borrero, E. E.; Escobedo, F. A. *J. Chem. Phys.* **2008**, *129*, 024115.
(73) West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104.
(74) Gabdoulline, R. R.; Wade, R. C. *Curr. Opin. Struct. Biol.* **2002**, *12*, 204.